Targeting Domestic Abuse by Mining Police Records

by

Matthew Paul Bland

Wolfson College

July 2019

This dissertation is submitted for the degree of Doctor of Philosophy

1 Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Matthew Paul Bland

July 2019

2 Abstract

Targeting Domestic Abuse by Mining Police Records

Matthew Bland

This dissertation presents findings from analyses of three large datasets of domestic abuse records sourced from multiple police forces in England and Wales. It seeks to address research questions in relation to the extent of repeat and serial abuse, concentration and escalation of harm, and the forecasting of future serious crimes. Using a variety of statistics, it shows that most victims and offenders report domestic abuse to the police forces just once in a multi-year period. Among these cases however, are many of the individuals who comprise very small 'power few' groups that account for most of total crime harm. Using the Cambridge Crime Harm Index as the instrument of measurement, analysis shows that 80% of cumulative harm is attributable to fewer than 3% of victims and offenders, and almost half of these most harmed victims or harmful offenders have only one record of domestic abuse in police databases. Police forces are therefore presented with a substantial challenge when it comes to preventing serious harm from domestic abuse, because in more than 40% of the most harmful cases they have not dealt with the victims or offenders of domestic abuse before.

Furthermore, among the victims and offenders who are linked to multiple records of domestic abuse, analysis detects no pattern of escalating severity. In fact, the first crime reported is, on average, the most harmful domestic crime reported to the police. This runs contrary to popular theories of escalation and further illustrates the forecasting challenge facing police agencies.

Contemporary harm reduction strategies have placed some emphasis on the management of serial perpetrators of abuse, but analysis shows that these do not offer a complete solution for harm reduction either. These analyses show that serial offenders account for only between 10% and 15% of all domestic offenders and contribute no more to the 'power few' than repeat offenders who have just one victim. However, analysis of the non-domestic crime records of domestic offenders does show that serial perpetrators are less specialised in domestic abuse than repeat or single-time offenders – they commit more non-domestic types of crime and account for more total crime harm overall. Serial and repeat offenders with the greatest generalist tendencies were shown to be attributed to the most

domestic abuse harm of any domestic offender types, indicating potential relevance to non-domestic offending records in the pursuit of predicting serious domestic crimes.

How then, might police agencies seek to identify the most harmful cases before they occur? This research explores a large bank of records relating to arrests for *any* type of crime, using a statistical model created by a supervised machine learning algorithm. This model processes each arrestee every time they enter custody and using predictors from 35 pieces of information primarily concerning the prior offending history of the arrestee, generates a forecast of future arrest for a domestic crime *within two years*. The forecast has three classifications – no arrest, an arrest for a 'less serious' domestic crime, or an arrest for a serious domestic crime. In this fashion, the model could, at best, predict 49% of future serious arrests. The other 51% of serious domestic crime arrestees have no prior arrest record in the two years preceding their serious domestic crime and so there is no opportunity to forecast their crime based on police records alone. However, within the 49% of serious domestic crimes that are possible to forecast, the model accurately predicts more than three quarters. Hypothetically then, using this method of forecasting the police could identify 37% of all future serious domestic arrests up to two years before they occur. This presents the foundation of a major opportunity to reduce the prevalence and harm of domestic abuse.

Taken together, these findings illustrate the scale of the challenge the police and other agencies face with reducing domestic abuse. A small proportion of individuals generate the majority of harm, but there are very limited opportunities to identify these individuals before the harm occurs. Yet, modern statistical techniques such as machine learning can help to target harm reduction strategies more precisely and even identify a sizeable proportion of serious cases before they occur. This dissertation presents with a series of ideas about how these objectives may be achieved and how the research can be developed further still.

3 Acknowledgements

These acknowledgements feel as though they were a long time in the making! In no particular order, I am especially grateful to the following people who supported me during this research.

Professor Lawrence Sherman and Dr Heather Strang – without their confidence, encouragement and support, I simply could not have started, let alone completed this work. I will always be grateful for the opportunities you have given me.

Dr Barak Ariel – not just an ace supervisor but a good friend and a good sport, especially in the face of 101 daft questions about statistical tests. (These are unlikely to stop).

Ann and Paul Bland (aka Mum and Dad) – who also gave me every possible opportunity and provided invaluable support in the form of childcare and occasional food packages. For the last few months of this thesis, they have been embroiled in a fight with cancer and just as they always have, showed me how to stick it out when things get tough.

Louise Bland – who was there beside me for the entire process, enduring many lonely evenings while I wrote or stared at the 'calculating cells' message on Excel. Thank you for listening and putting up with me.

Jacob and Sophie Bland and Luke and Chloe Sadler – my wonderful children. I will never forget attempting to explain random forests to an 8-year old Jake or having to undo the many typing forays of a 1-year old Sophie. I hope you will all be impressed by the final result.

4 Summary

This research considers what insights into domestic abuse may be gained from analysis of large police datasets. Rising demand from domestic abuse cases is stretching police forces in England and Wales, not least because of a resource intensive minimum response standard that includes police officer attendance at all calls and extensive risk assessment of every case. Yet many of the features of this response are not founded on rigorous evidence. While the evidence-base grows, the police service must look at is own resources for answers about how to refine its responses. Its own sizeable databases, which represent the largest pool of domestic abuse records of any single profession, are one possible option.

Since 2016, police forces have been required by law to record domestic abuse crimes in a standardised way, enabling better cross-jurisdictional analyses. Even before this point, there was consistency in the way many police agencies kept domestic abuse records and the evolution of these data presents an opportunity to explore whether the policing response may be more precisely targeted toward the highest harm cases. This research uses a combination of descriptive and inferential analytical techniques on three large datasets provided by various police forces in England and Wales. The following summary outlines the main findings.

To what extent is domestic abuse a repeat phenomenon?

Despite common perceptions that it is most frequently a repeat occurrence, in fact police forces only record one case of domestic abuse for around three quarters of victims and the same proportion for offenders. However, once additional crimes are reported it becomes increasingly probable that further crime records will follow. Broadly, this likelihood increases with each additional consecutive crime record and becomes ‘probable’ (i.e. more likely than not likely) after the third crime report.

In the cases that do report repeated domestic abuse, is there evidence of escalating severity over time?

No. In fact the opposite is true. There is strong evidence that among both victims and offenders with enough repeat cases to provide a window of opportunity for escalation to be observed (defined in this analysis as five or more domestic abuse reports), that the first recorded incident is, on average, the most serious crime they report to the police.

How do you determine what types of crime are more serious?

There are a number of crime severity or crime harm measurement instruments available to researchers now. The one best suited to the purposes of this analysis was the Cambridge Crime Harm Index (CCHI), which rates more strongly than other options in terms of reliability, practicality and legitimacy. It weights crime classifications by the minimum number of days in prison a judge or magistrate is recommended to apply for first-time offenders. For example, a murder is weighted at more than 5,000 days, whereas a common assault without injury is weighted at 10 days.

What does the use of the CCHI tell us about how harm is distributed?

It tells us that harm is spread among both victims and offenders in an extremely disproportionate way. If we rank order offenders in descending order of total harm, it shows us that 80% of all harm is linked to fewer than 3% of offenders. The same is also true for victims.

Who makes up these 3%s?

The 3%s (or 'the power few') are victims and offenders linked to at least one crime that is at least equivalent to the minimum sentence for grievous bodily harm without intent (545 days). Typically, victims and offenders in their respective 'power few' are older and less frequently male than victims and offenders not in the 'power few'. Most of these individuals are repeat cases, meaning there is some window for prediction because there may be at least one crime that occurs before the serious crime. Hypothetically, police might be able to spot something in those precursor crimes that predicts a future serious crime but around 40% are cases in which the victim or offender is known to police for only one crime – the one that was high harm. This makes prevention a difficult proposition for police forces – how do you stop these harmful crimes from happening when you have no prior indication of the participants?

What about serial cases – are these a key part of the 'power few'?

Of those victims and offenders who are linked to repeat cases of domestic abuse, just less than half have more than one unique 'other party' involved (the definition of 'serial'). Police forces have been primarily interested in serial perpetrators on the understanding that they are more harmful than domestic offenders who commit crime against just one victim, but this

premise is not wholly supported by this research. Serial perpetrators are not statistically different from repeat offenders (those who are linked to only one victim) or single-time offenders (those linked to just one crime) in terms of ethnicity or gender or age. More pertinently, serial offenders commit a different mix of domestic abuse than their counterparts, with fewer violence with injury crimes, fewer rapes, but more criminal damage and more of other forms of 'less serious' domestic abuse, such as theft and fraud. Analysis of CCHI shows that serial perpetrators are no more harmful than repeat offenders, and while they are proportionally more prevalent among the most harmful group of offenders than they are among the offending population in general, the 'power few' is still largely made up of offenders with just one domestic abuse crime record. However, serial perpetrators do commit a greater number of non-domestic abuse crimes and generalist serial offenders – those who commit multiple types of violent and non-violent crimes are, together with generalist repeat offenders, the most harmful of all types of domestic offender.

If serial perpetrators are only a small part of the 'power few', how else can police forces target offenders in this group?

Whatever methods police forces choose for targeting offenders, it is likely that a major problem will remain: how to identify high harm cases when there is no or little prior record of domestic abuse? Current risk assessment processes are predominantly focussed on risk assessing cases *after* they have reported a domestic incident, meaning that a large proportion of the most harmful crimes will not be exposed to the large police and partner infrastructure dedicated to preventing harmful domestic abuse until it is too late.

One option is to leverage arrest records for *all* types of crime to see if prior arresting history for non-domestic crimes is predictive of future domestic behaviour. These records could be processed with a classification-based algorithm such as 'random forests'. Using a dataset from one police force, the random forest statistical model we developed could 'reach' 49% of serious domestic crimes. The model was able to identify more than three quarters of those (37% of all serious crimes). In other words – by screening all arrestees using this model, police could identify 37% of all future serious domestic crimes, regardless of the prior domestic history of the victim or offender.

How does the algorithm work?

The algorithm uses the advent of an arrest (for any type of crime) to trigger a forecast. It then processes the arrestee based on more than 30 variables, predominantly concerning the

arrestee's prior offending record. It produces a forecast for future behaviour within the next two years – (a) no domestic arrest, (b) an arrest for a 'less serious' domestic crime or (c) an arrest for a serious domestic crime. The forecast is derived from hundreds of decision trees calibrated using tens of thousands of arrest records.

Could such a process work in practice?

Yes. Providing a police force can leverage its data with the appropriate statistical expertise and computer processing power, it is possible to deploy an application in custody suites to forecast the subjects of all arrests. Any agency wishing to use such a method would need to decide whether the forecasting error rates were acceptable. Not all serious domestic crimes involve an offender with a prior arrest of any kind, so the algorithm cannot predict such cases (although neither can the existing risk assessment procedures used by police forces). The algorithm could at most predict around half of future serious abuse, and within that proportion it accurately identifies 77%. However, to achieve that level of accuracy it is quite cautious – nine in every 10 forecasts of serious abuse would be incorrect – a high rate of 'false positives', but again, an improvement on the current system. Crucially though, the model is almost always correct when making forecasts of *no* future abuse. The proportion of 'no abuse' forecasts that are subsequently arrested for a serious domestic crime is less than 0.1%

The algorithm may be adjusted to suit different agency preferences such as capacity to process 'high risk' cases and it could almost certainly be improved with the addition of new predictor variables and further testing. It does not specify what agencies should do with individuals who are forecast as likely to commit a serious domestic crime, and the forecast in itself is not a method of prevention, merely provides the opportunity for an intervention to be more precisely applied.

What are the implications of these findings for police agencies and domestic abuse researchers?

The main implications from the findings are best summarised as follows:

1. Theories of universal escalation should be revised to reflect that the phenomenon does not occur in police records. This does not mean that escalation does not occur, but it does indicate that if it does, the police are largely blind to it.

2. Future research should focus on the issue of desistance, which is a dominant characteristic of police domestic abuse records. It may be that cases desist in the true sense (no further crime, perhaps because of separation or estrangement) or that abuse simply goes 'underground'. It is vital that police and researchers understand the extent of both and the reasons.

3. Serial perpetrators are important to offender management strategies, but such strategies should not be based solely on this group. Perpetrator management should focus at least equally on repeat offenders who commit crimes against a single victim. A register of serial perpetrators would not account for the majority of recorded domestic abuse harm.

4. Examining an offender's non-domestic abuse history presents offender managers with a relatively simple and viable way of prioritising their efforts to target the potentially highest harm offenders.

5. Generally, domestic abuse prevention strategies should recognise that the most harmful cases and individual offenders are proportionately few. Future research and practice should focus on establishing what characteristics definitively separate 'the power few' from the general population of domestic victims and offenders. This need not suggest that the general population of cases is unimportant but should enable a differential response to evolve on a more systematic basis. In the context of increasing demand and declining resources in policing, such an approach is inevitable.

6. Actuarial forecasting tools are promising and there is much scope for researchers and practitioners to develop them further but careful assessment of the political ramifications must be closely considered to ensure responsible and legitimate use of actuarial models.

5 Contents

6 List of Tables

Question Number	Question	Findings

7 List of Figures

8 Introduction

8.1 Research overview

This research explores what domestic abuse records kept by police forces can tell us that may assist in refining harm reduction strategies employed the police. Domestic abuse has emerged as a priority in policing, particularly in the last decade. There is extensive evidence that this form of crime is a matter of grave concern to public health and safety in the twenty-first century – it is widespread, expensive and a major drain on policing resources. Official statistics estimate almost two million adult victims per year, a prevalence of 6% of all adults (ONS, 2017) and the most recent comprehensive assessment of financial cost, albeit over a decade old, (Walby, 2009) placed the cost in the billions of pounds to service providers, employers and victims. With economic inflation and rising crime levels since Walby's estimate, it is likely domestic abuse costs the public purse even more today. Furthermore, domestic abuse is a major factor in the most serious crimes - domestic circumstances feature in a third of murders in England and Wales and in more than a tenth of all crimes recorded by the police (ONS, 2018).

The purpose of this work is to contribute to the evidence base for tackling domestic abuse. This has greatly expanded in recent years, in which research has delved deeper into the impacts of particular subcategories of domestic abuse. Substantially more is now known about the impacts of forced marriage (Watts and Zimmerman, 2002), revenge pornography (Henry and Powell, 2014; Bond and Tyrell, 2018) and financial abuse (Sharp-Jeffs, 2015, 2017) than at any point in the past. This body of research has developed against a background of an emerging evidence-based policing (EBP) movement in England and Wales. Led by partnerships between the police professional body, the College of Policing (CoP), the National Police Chiefs' Council (NPCC) and academic institutions, EBP forms a central tenet of modern policing with the aim of improving practice through the accumulation of robust empirical evidence (Lum and Koper, 2017; Neyroud and Weisburd, 2014). The nature of that empirical evidence has been the subject of much debate (see Cockbain and Knutsson, 2014; Sparrow, 2011; Weisburd and Neyroud, 2013). Sherman, who first established the term 'evidence-based policing' (Sherman, 1998), has proposed a framework for viewing EBP activities through the lenses of targeting, testing and tracking (Sherman, 2013), and appraised the development of domestic abuse evidence through this framework (Sherman, 2018). It is specifically in this context that this research is positioned, building on recent findings in

targeting evidence (Barnham, Barnes and Sherman, 2017; Bland and Ariel, 2015; Bridger, Strang, Parkinson and Sherman, 2017; Chalkley and Strang, 2017; Kerr, White and Strang, 2017; Sherman and Strang, 1996; Thornton, 2017) and attempting to realise the promise of new analytic techniques and sources applied in criminology, such as:

- The availability of 'big data' in the manner described by Sherman (2018) and demonstrated by previous work in other fields of criminology (Berk, Sherman, Barnes, Kurtz and Ahlman, 2009).

- The development of new harm measurement instruments such as the Crime Severity Score (ONS, 2016b), the Cambridge Crime Harm Index (Sherman, Neyroud and Neyroud, 2016) and the updated Home Office Cost of Crime Estimates (Heeks et al., 2018).

- The potential application of new machine-learning algorithms to big datasets, such as has been demonstrated in several recent publications (Berk, 2012; Berk, Sorenson and Barnes, 2012).

Although the evidence base for domestic abuse is already comparatively rich, at least in the context of the general depth of rigorous evidence on policing activities, there has been long been an ongoing demand from practising agencies to acquire information and evidence that can further shape strategy (see for example Shepherd, 1998; Sherman, 1992, 2018). The impact of 'austerity' in the United Kingdom was to reduce the capacity of all government sectors, including those with primary responsibility for dealing with domestic abuse (Neyroud, 2015). Yet at the same time, scrutiny from the national police oversight body, Her Majesty's Inspectorate of Constabulary, Fire and Rescue Services (HMICFRS), has aimed specifically at domestic abuse and has been highly critical of the police response (HMICFRS, 2014a). As a consequence of this, as well as a separate critical review of crime recording practices (HMICFRS, 2014b), police forces have attached greater priority to identifying and responding to domestic reports, resulting in recorded crime numbers rising steeply at a time when other sources showed a decline in the prevalence of self-reported domestic abuse in the adult population (ONS, 2017).

With around 20% fewer police officers than in 2010 (ONS, 2019), responding to the additional demand has been a challenge for police forces. Arrests and charges have declined (Ariel and Bland, 2019; ONS, 2017, 2018), and the inspectorate continues to highlight deficiencies in the police response in areas such as identification of risk (HMICFRS, 2017, 2019). This context provides an opportunity for well-developed targeting research to help shape domestic abuse strategies, and it is precisely this opportunity which this research aims to address.

There remain questions about the extent to which the evidence influences what the police (or other responding agencies) actually *do* in practice. As we will explore further, much of the current response to domestic abuse is not driven by evidence, and the collection of data concerning victims and offenders remains, to this point, an under-utilised resource in the development of domestic abuse strategies. It is this area that is our target. The overarching aim of this research is to add to the existing evidence base in ways that could usefully contribute to front-line strategies and underpinning theories, by exploiting the potential of an existing resource abundant in every police force - crime data.

In this respect, we gathered hundreds of thousands of anonymised domestic abuse records from police forces around the country. These records related to crimes, arrests, offenders and victims, and all of these data resemble the typical sorts of information every police agency has ready access to. These records were assembled into three large datasets and analysed using a variety of statistical procedures, ranging from the very simple (rates and proportions) to the very complex (a machine learning algorithm). Each procedure that has been used has been selected with a particular research aim in mind and these aims were selected because they represent issues of high relevance to practitioners and researchers, and because there are gaps or uncertainties in what these groups know about these issues.

Throughout these issues the topic of 'harm' is a persistent feature. As we will explore, much of the current response to domestic abuse is geared towards the identification of 'high-risk' cases and subsequent action to negate that risk. It is logical to argue that 'risk reduction' is actually an outcome that the police service and its partners are seeking, but this leads to the inevitable next question: the risk of what? The answer seems perfectly logical – serious harm to the victim – but this in turn raises a difficult question for a crime researcher. How does one measure harm? Harm is a subjective concept, particularly among practitioners; what constitutes serious harm for one person does not necessarily do so for another, and in this

spirit a number of harm measurement tools have been developed. However, there are currently no national guidelines to guide this debate in any particular direction. This is the first challenge that this research seeks to overcome: the selection of an appropriate instrument for the measurement and tracking of harm to facilitate the further exploration of police data.

Armed with an appropriate tool to measure and track harm, we return to the key research questions. These are organised into five principal categories: repeat abuse, serial abuse, escalation, concentration of harm and forecasting. Each category has its own distinct questions of interest which we will use the data and statistical procedures to attempt to answer. These questions are as follows:

Repeat Abuse[1]

1. What is the prevalence and extent of repeat victimisation of domestic abuse?
2. What is the conditional probability of further domestic abuse associated with each consecutive victimisation?
3. What is the prevalence and extent of repeat offending of domestic abuse?
4. What is the conditional probability of further domestic abuse associated with each consecutive offence?

Serial Abuse[2]

1. What is the prevalence and extent of serial abuse among victims of domestic abuse?
2. What is the prevalence and extent of serial abuse among offenders of domestic abuse?
3. Are serial perpetrators demographically different from repeat offenders[3] or single-time offenders?
4. What types of domestic abuse crime do serial perpetrators commit and how harmful are they?
5. Do serial offenders cause more domestic abuse harm than repeat or single-time domestic offenders?
6. To what extent do domestic abuse serial perpetrators commit other forms of crime, and how does this compare with repeat or single-time domestic offenders?

[1] Repeat abuse is defined as multiple domestic crimes regardless of the identity of the other party involved
[2] Serial abuse is defined as an offender with multiple different victims, or a victim with multiple different offenders
[3] Repeat offenders in this sense are those which offend multiple times against just one victim

Escalation

7. Is there evidence of escalating harm in each consecutive domestic victimisation?
8. Is there evidence of escalating harm in each consecutive domestic offence committed?

Concentration of Harm

9. What is the extent of concentration of harm among victims of domestic abuse?
10. What is the extent of concentration of harm among offenders of domestic abuse?
11. To what extent do the police have prior knowledge of the group of victims suffering the most harm?
12. To what extent do the police have prior knowledge of the group of offenders committing the most harm?

Forecasting

13. What proportion of all arrestees go on to commit domestic abuse within two years?
14. What proportion of serious domestic abuse arrestees have prior records for domestic abuse?
15. Can antecedent inputs predict future serious domestic abuse cases to a high degree of accuracy?
16. If so, which inputs have the greatest impact on accuracy?

8.2 Dissertation structure

These questions cover a lot of ground, so this thesis is organised into discrete chapters for ease of reference. The initial chapters introduce the context of the research in greater detail. Chapter 9 presents the current circumstances relating to domestic abuse in England and Wales. It defines key terms, explains aspects of the police and partner responses and summarises the main challenges.

Chapter 10 reviews how domestic abuse is measured and includes an assessment of the suitability of police records for this type of research.

Chapter 11 considers the issue of harm and how to measure it. It begins with a discussion of how harm is defined and then presents an appraisal of the various harm measurement instruments available, before concluding which to use.

The final contextual chapter is a long one. Chapter 12 examines the existing literature in respect of the key issues this research examines. This chapter is sub-divided into the principal research themes.

The next set of chapters present the main body of this research. Chapter 13 presents the research questions again in more detail. Chapter 14 sets out the methodology used to address each of those questions. Chapters 15 to 19 detail the findings of these procedures. Each of these chapters is dedicated to one of the five principal research themes.

In Chapter 20, we discuss the implications of the findings. This chapter summarises the overall themes and examines the 'so what?' aspects of policy, research and theory. It also sets out the primary limitations of each analysis. Chapter 21 develops the themes further and presents concluding thoughts, making references to the contextual points set out in Chapters 9 and 12 in particular.

Finally, there is an appendix included for readers wishing to consider the more technical aspects of how the forecasting model presented in Chapter 19 was developed.

9 Domestic Abuse in England and Wales

9.1 Chapter roadmap

This chapter sets out the operating context in which this research has taken place. It begins with a description of the definition of domestic abuse, explaining its principal characteristics and the nature of the problems it poses to society. The main body of the chapter presents a summary of the different aspects of the police response to domestic abuse, which frequently involve partner agencies too. Though non-police responses are not our primary focus, examples of partner activities are highlighted throughout this chapter.

9.2 How domestic abuse is defined

This work uses the standard UK cross-government definition of domestic violence and abuse, reprinted here for clarity:

> Any incident of controlling, coercive, threatening behaviour, violence or abuse between those aged 16 or over who are, or have been intimate partners or family members regardless of gender or sexuality. The abuse can encompass, but is not limited to psychological, physical, sexual, financial or emotional. (Home Office, 2012)

The definition includes the elements that were added in 2012 to encompass coercive and controlling behaviour and 16–17-year-old victims, following a public consultation. This replaced the previous separate definitions published by the Home Office and the Association of Chief Police Officers. The definition change was accompanied by the introduction of a new criminal offence in relation to coercive and controlling behaviour, a concept first established by the criminologist Evan Stark (2007). Coercive and controlling behaviour is currently the only domestic abuse–specific offence in British law; in all other cases, domestic abuse is effectively a circumstance attached to another legally defined criminal offence, for example, assault, rape or burglary. As such, accurate recording of abuse is a complex issue and separate from the usual form of official crime counting.

In practice, the police are the service with primary statutory responsibility for determining and recording a criminal act, but it is by no means the only agency that has contact with domestic abuse victims. Schools, hospitals, charities, doctors' surgeries, housing providers, social workers and more all come into contact with domestic abuse cases and are required to apply the cross-government definition. Health practitioners in particular, have a

critical role in identifying and referring victims to specialist services. But it is only the police service that is administratively required by the Office for National Statistics to record the occurrence of domestic abuse in official records. Despite this, the cross-government definition obviously offers the most practicable definition for research, and as we stated at the outset, police records are the only data source used in this research.

Other government services such as the National Health Service provide employees with guidance on how to identify domestic violence and abuse which is consistent with the national definition, but there is no cross-agency or nationwide requirement to document domestic abuse. In these settings, guidance commonly recommends caution and the seeking of victim consent to record (see domesticviolencelondon.nhs.uk). The only other 'regulated' sources of domestic abuse records are the Crime Survey of England and Wales, which comprises interviews and self-completion questionnaires, and criminal justice agency statistics, which are connected to police records (having been passed into the criminal justice system by the police service). The domestic abuse charity SafeLives maintains a national database of high-risk cases which includes some that are not in police datasets, but in all examples these datasets fall within the scope of the national cross-government definition.

While this definition is most relevant to agencies in England and Wales, it is not without relevance to researchers and practitioners in other countries. There may be no internationally accepted definition of domestic abuse, but intimate partner violence is a well-known and researched subject. The principal difference between the UK cross-government definition used here and the typical description of intimate partner violence is the former's inclusion of family members over the age of 16. This form of abuse does not make up the majority of domestic abuse as recorded in England and Wales (ONS, 2018), meaning practitioners and researchers of intimate partner violence may still derive some broad meaning from the definition we have used in this work.

9.3 Characteristics and terms used

How the subject matter is defined is important, as does how much of it there is (or at least, how much is known about), and what the main principles of current research and practitioner approaches thereto are. This last issue is considered in this chapter and the matter of how much domestic abuse there is, is considered in Chapter 11, but prior to their consideration it is worthy of our time to expand on some of the more basic characteristics of domestic abuse, for the sake of clarity and introduction to readers new to the subject.

First and foremost, domestic abuse is a form of crime with several names. It is often referred to as domestic violence or intimate partner violence. It is less frequently known as wife-battering or spousal abuse, which were popular terms in the last century. Neither term is invalid, but neither represents the full spectrum of relationship circumstances in which a domestic crime can occur. Domestic abuse as we consider it in this research, concerns personal crimes between intimate partners or family members above the age of sexual consent. These crimes are not necessarily violent. Hypothetically, any form of crime could be domestic-related, although the majority recorded by police are usually classified as violence with or without injury (ONS, 2017, 2018).

Each domestic crime recorded in England and Wales has consistent components which we will refer to throughout this research. For clarity, the most commonly used are explained as follows:

Victim: the individual against whom the crime is committed. In this research, the term 'survivor' is also covered by use of this word.

Offender: the individual who is accused of or proved to have committed the crime. This research makes no distinction between the judicial status of a case so all suspects are described as 'offenders', and the term 'perpetrator' is often used interchangeably but always with the same definition.

Dyad: the unique combination of a particular victim and particular offender. This research does not use dyads as an analytical unit but reference is made to the term in the analysis of literature and discussion of implications.

Crime Classification: Each case is assigned a Home Office classification based on its nature. For example: burglary, assault, grievous bodily harm etc. Technically, any form of crime can be domestic abuse, but domestic abuse is *not* a crime classification in its own right.

Repeat Abuse: Although some agencies specify a minimum level of occurrences, or a minimum or maximum window of time, in this research any multiple instance of domestic abuse between the same dyad is referred to as 'repeat abuse'.

Serial Abuse: When an offender or victim is attributed to multiple domestic abuse crimes but within multiple dyads, this research refers to these cases as 'serial abuse' rather than repeat abuse. There has been considerable interest in the last decade in understanding and targeting offenders of serial abuse in particular, which we will examine further.

Domestic Non-Crimes: Also known as domestic incidents, these are calls for service that the police attend and determine to be relating to a domestic dispute, but which do not meet any threshold or criteria for a criminal act. As such no victim or offender status is assigned (though the details of the individuals involved are kept).

9.4 The main problems facing responders

There can be little argument that domestic abuse is a 'just cause' for a researcher's attention. Even before one considers the plethora of official statistics on prevalence, harm and cost, there is the moral heart of the matter: that people ought to be at liberty to have personal lives free from crime or abuse and it is our duty to investigate situations where this is not the case. As Stark (2007) explains, along with practitioners, researchers share the burden of formulating effective responses, and as already outlined, this is precisely the spirit in which this research has been designed. However, before we can establish a meaningful research plan, we must understand the nature of the problem(s) with domestic abuse so as to identify the areas of most pressing interest. As a first step, this section attempts to outline the nature of the domestic abuse problem in England and Wales as a frame of reference for the specific questions this research addresses.

It must be emphasised from the outset that defining strategic domestic abuse problems is itself a problem. It is universally acknowledged that domestic abuse is underreported (see Brimicombe, 2018) – which is to say more (much more) takes place than is recorded in official records (see ONS, 2018). The main source of official records on domestic abuse in England and Wales is those kept by the police, both in the forms of recorded crimes and domestic 'non-crime' incidents. The numbers of official records increased substantially in response to the police inspectorate's domestic abuse reports in 2014 and 2016 and parallel pressure on crime recording standards. Commentators also speculate that the rising media profile and service prioritisation of domestic abuse has led to more victims reporting to police (ONS, 2017), but no robust evidence yet supports this hypothesis.

However, although the rise in recorded non-crimes and crimes represented 88% between 2016 and 2018 (HMICFRS, 2019), the extent to which police records reflect prevalence is still claimed to be low. The charity Women's Aid conducted a census survey of more than 12,000 domestic abuse community-based services users and 2,000 women's refuge users in the UK and found that just 28% and 43%, respectively, were reporting to the police (Women's Aid, 2017). These results may be limited – the survey period utilised a 'day to

count' and 'week to count' structure and had an overall response rate of just over 50%, but even if extrapolation is disregarded, the emerging trends were notable in their own right.

The specialist UK domestic abuse charity SafeLives also eschews the use of official statistics (SafeLives, 2018) and instead uses its own survey-based dataset of more than 35,000 records, claimed to be the 'largest database of domestic abuse cases in the UK'. Despite this, both Women's Aid and SafeLives quote the Crime Survey of England and Wales as the main source of statistics when presenting information on prevalence.

Herein lies the most fundamental problem: what actually are the problems? With such disagreement about an ingredient as essential to defining problems as source data, it is perhaps unsurprising that there is ambiguity, disagreement and confusion around domestic abuse strategies. The roots of this most fundamental issue seem to be intrinsic. Domestic abuse is a form of interpersonal crime that takes place in an intimate environment, frequently a home, unseen by witnesses. The suffering and perpetrating parties have a relationship which predates and will likely post-date the offence. Victims often live in fear of reprisals for themselves, their children or their pets. Offending is often subtle to the point of being intangible, as Stark (2007) eloquently demonstrated in his descriptions of coercive control. If all of these obstacles were not enough, there is still the matter of the lack of a single recording environment for a victim's case but this is, to a larger extent, a matter of necessity. Victims should not be compelled to report to the police, nor have their personal data shared outside of the organisation they have chosen to approach. It is not merely a question of whether a single data source could be created for domestic abuse; there is a very important question of whether one ought to be created in the first place.

The problems then are perhaps best described as 'basic' in nature. For police, rising demand is a major problem which gives rise to a primary concern: how to provide a high-quality response to each case. Given the shrinking capacity of the police in recent years (ONS, 2019) this is a dilemma without an immediately feasible solution. Instead then, the issue becomes about how to best manage that demand. Here, there are three key aspects:

1. Triage: how can the police identify which cases most need their limited resources?
2. Differential response: what responses work most effectively with whom?
3. Prevention: how can the police (or other agencies) prevent domestic abuse from occurring in the first place, and thereby reduce demand and free capacity for response and further prevention.

This research is concerned with two of these three aspects. We exclude any substantial consideration of differential responses, which require individual evaluations. Our research questions are instead aimed at primarily contributing to evidence relevant to the first and third aspects of the police problem: how might police forces triage cases and how they might prevent them? We will examine these issues in more detail in later chapters. Before doing so, it is important to consider how the police service typically responds to domestic abuse at present.

9.5 Police responses to domestic abuse

To further contextualise these problems and issues this section sets out a brief description of the current landscape for domestic abuse responses in England and Wales at a strategic level. Each description briefly reviews the evidence that does (or does not) underpin each activity. The intent here is not to provide a comprehensive critique of current strategy and its supporting evidence so much as to further frame the relevance of the research questions this work seeks to address.

9.5.1 Initial response: Mandatory police attendance

In England and Wales, the policing response to domestic abuse is characterised by the phrase 'positive action', which encompasses both the initial response to a call to police and the actions taken once a police officer arrives on the scene. 'Positive action' is derived from the 'positive obligation' component of the Human Rights Act and is translated in the Authorised Professional Practice for policing (College of Policing, 2018) as necessitating the attendance of a police officer in each and every domestic abuse case. Call takers must grade domestic abuse calls for an immediate response and attending officers must use the police 'national decision model' to assess whether immediate action is required once they arrive.

Latterly, the practice of wearing and operating a body–worn video camera has been promoted as effective in increasing the proportion of cases resulting in charges (Owens, Mann and McKenna, 2015). HMICFRS found that the rate at which police forces comply with this policy is 'improving' but has a long way to go (HMICFRS, 2017, 2019). Given this verdict from the inspectorate as well as the current volumes and recent trends in domestic abuse records held by the police, it is fair to say that this policy places a substantial demand on resources.

The College of Policing conducted a systematic review of research evidence on the effects of police attendance at domestic abuse events in 2016. The review synthesised evidence from nine studies, which were mostly from the USA and not from the recent past. It concluded that there was very little evidence that police attendance had any impact. The one exception to this conclusion came from Felson, Ackerman and Gallagher (2005), which found no evidence of retaliatory violence in response to a report made to the police.

9.5.2 Arrest policy

England and Wales police forces do not operate a statutory mandatory arrest policy; it is in fact impossible to statutorily compel a police officer in these countries to arrest a person. The term 'mandatory arrest' is associated with the move made by many US states following the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984), but is often confused with the 'positive action' policy operated in England and Wales. Authorised Professional Practice (College of Policing, 2018) states that officers must justify any decision not to arrest, as part of a suite of actions to make victims safe. These actions include other police powers, such as the issuing of a civil order or a caution, but each of these has particular policy specifications, considered below. The fact is that arrest does not take place in the majority of domestic abuse events and has declined proportionally in recent times (ONS, 2018). This trend has drawn criticism from the police inspectorate and been further complicated by changes to police bail powers (HMICFRS, 2019).

The effects of arrest on domestic abuse recidivism are one of the best researched areas in the field (see Vigurs et al., 2016 for a summary), but the results are not uniform. Sherman and Berk's initial randomised experiment in Minneapolis (1984) concluded lower levels of recidivism in cases assigned to arrest than in those assigned to 'advice' and was the catalyst for widespread state legislative changes to domestic violence arrest policies in the USA as well as a slew of replication studies in the form of the Spouse Assault Replication Program (SARP). However, Sherman, Schmidt and Rogen (1992) found mixed results in other sites, and Maxwell, Garner and Fagan (2002) found moderate effects on prevention in police records but statistically significant reductions in victim reports. All in all, a mixed bag.

Other studies also contributed to the contradictory evidence. Cho and Wilke (2010) examined the effects of arrest on repeat victimisation among males and found no deterrent effect, and both they and Felson, Ackerman and Gallagher (2005) hypothesised that simple attendance by the police had as much impact on recidivism as arrest. Others have found

outright detrimental effects from arrest. Iyengar (2009) compared the rate of domestic murders between US states with and without presumptive or mandatory arrest policies and found that those states with presumptive arrest policies saw a greater rise than those without. Iyengar speculated that this was due to arrests undesired by victims having a suppressing effect on future reporting and a 'reprisal effect' from the arrest policy, but neither theory has been tested sufficiently to establish a causal effect. Sherman and Harris (2015) were able to draw stronger causal links between arrest and increased victim mortality, albeit general and not from homicide. Following up on participants of the original Sherman and Berk Minneapolis experiment from 1984, they found that victims whose partners had been subject to arrest instead of advice had a 64% greater chance of having died in the subsequent 25 years. This was particularly evident among employed African Americans and not influenced by homicide (only three of 91 deaths were homicides).

Myhill (2018) argues that the absence of definitive evidence of the positive effect of arrest on recidivism does not mean it is an inappropriate measure because it enables more comprehensive recording to shed light on patterns of coercive control. The introduction of this form of criminal offence, Myhill argues, means that dealing with domestic abuse is a distinctly different premise for police officers in England and Wales than in, say, the USA, thus negating calls by others (Sherman, 2015) to explore alternatives. This remains the prevailing view at the time of writing, with the police inspectorate continuing to highlight falling arrest rates in police forces as a cause for concern (HMICFRS, 2019).

9.5.3 Alternatives to prosecution

9.5.3.1 Civil orders

Police officers may apply for civil orders in cases where an arrest has not been made (although they are often sought when an arrest has been made). Such orders, known as Domestic Violence Protection Orders (DVPO), can be applied for without victim support, and are authorised or denied by a magistrate. DVPOs are preceded by notices issued by officers while the order is being prepared (DVPNs). These require the authorisation of a Superintendent and challenge the capacity of that rank. The police inspectorate reported in 2017 (HMICFRS, 2017) that it was increasingly concerned with the declining use of DVPNs and DVPOs, particularly in light of an evaluation conducted in 2013, when the scheme was being piloted (Kelly et al., 2013), which found that the tactic was 'associated with reduced levels of re-victimisation'. This seems to be a highly contestable claim. Kelly et al.'s study

was a case-matched sample design which, once filtered for prior domestic crime history, left a sample size of just 123. Among this cohort, a statistically significant reduction of one domestic crime was observed. For cases reporting for the first time, there was no effect. Despite the modesty of these findings, they were the catalyst for DVPN/Os being implemented across the country. There has been only one significant study of this tactic since. Smith (2016) conducted a case-control analysis of DVPN/Os issued in Hertfordshire and found no significant differences between the DVPN/O group and the matched control sample in domestic crime before and after the issue of the order.

9.5.3.2 Cautions

The College of Policing guidance to police forces in England and Wales is emphatic about cautions, stating that they are rarely appropriate in domestic abuse cases, and that for intimate cases a conditional caution[4] would likely never be appropriate. Despite this, Westmarland, Johnson and McGlynn (2018) found that many forces were using out-of-court resolutions such as cautions on a regular basis. A recent Ministry of Justice evaluation of out-of-court criminal justice resolutions was inconclusive about their use in domestic abuse cases because of the absence of a counterfactual but found no significant difference in reoffending among domestic offenders in the pilot areas at a three-month review point (Ames et al., 2018).

The prevailing view about cautions being unsuitable was strongly challenged by a recent experiment in Hampshire (Strang et al., 2017), which tested the effects on crime count and harm among a cohort of 154 male domestic abuse offenders who were compelled under conditional caution to attend two day-long workshops. In comparison to the randomly assigned control group, the workshop attendees were re-arrested for domestic abuse 21% less often and with 38% less harm (as measured using the Cambridge Crime Harm Index[5]). At the time of writing, several forces in the country were planning to embark on their own pilots of conditional caution workshop schemes for low-risk offenders.

9.5.3.3 Restorative justice

Professional practice advice is equally clear that restorative justice tactics are as inappropriate in domestic abuse cases as cautions, yet here too there is emerging evidence that challenges this position. Ptacek (2017) best summarised this by highlighting the promising evidence on

[4] A conditional caution is a classification of investigative outcome in England and Wales which implies that a caution will not be issued providing specified conditions are met.
[5] The Cambridge Crime Harm Index is explained in Chapter 3

the efficacy in crime reduction of restorative approaches collated by Strang et al. (2013) and setting it against a lack of rigorous evidence in any direction as far as domestic abuse or intimate partner violence is concerned, with control groups used only in Pennell and Burford (2000) and Mills, Barocas and Ariel (2013). The result is that the research field neither knows whether victim–offender conferences can be effective in domestic cases, nor has any evidence to the contrary.

9.5.4 Risk assessment

Risk assessment is a central tenet of the domestic abuse policing strategy in England and Wales. With risk assessment defined as a cyclical process of estimating 'the likelihood and nature of a risk posed by a perpetrator to a particular victim, children or others' (College of Policing, 2018), the police policy thereon is to require each attending officer to conduct a structured professional judgement exercise against a series of predetermined questions posed to the victim. This process has been in place since 2008, when the Home Office and Association of Chief Police Officers endorsed the risk assessment model known as Domestic Abuse, Stalking and Honour-Based Violence, or 'the DASH', championed by the charity Co-ordinated Action Against Domestic Abuse (now SafeLives). The DASH was constructed based on the 'SPECS+' model used in London but was not evaluated or tested in any structured way prior to its implementation (Myhill, 2018).

The College of Policing (2018) asserts that evidence concerning domestic violence predictors is limited but lists 10 predictors based on professional expertise. This list includes (inter alia) previous physical assault by the perpetrator, escalation, animal abuse, child abuse and suicidal tendencies, all of which are reflected in the DASH. Though the DASH emerged from Richards' (2006) study of domestic homicide cases, it is not entirely clear how its predictor elements took shape, and the tool has been defended as preventative rather than predictive in nature (Richards, Letchford and Stratton, 2008). Regardless, the DASH has been subject to a range of critical studies in recent years. Thornton (2017) found the DASH to have low predictive validity in relation to domestic homicide and 'near-miss' cases, the majority of which the police had no prior contact for, and a very high false positive rate (in which high risk cases did not result in a deadly crime). Strang and Chalkley (2017) replicated Thornton's work and found a false negative rate of 67%. They went on to highlight suicide and self-harm warnings on the part of male perpetrators as having high predictive validity (suicide is specifically mentioned in DASH). Robinson (2016) also found that the DASH was not being used consistently in all police forces, and in 2017, the College of Policing undertook to

review the DASH question set, subsequently amending it to place greater emphasis on coercive and controlling behaviour. In 2019, Turner, Medina and Brown (2019) found the tool to be underperforming, little better at prediction than chance and with every question to be weak predictors of future abuse, at best. Even ignoring other studies, this paper alone casts major doubt on how fit for purpose the DASH is as a risk assessment tool and highlights the acute need for improvement.

9.5.5 Advocates

In England and Wales, Independent Domestic Violence Advisors (IDVAs) operate independently of police forces but are sometimes located alongside domestic violence units. It is normal practice for all cases categorised as high risk by the DASH process to be assigned to an IDVA, whose role it is to support the victim by acting as an advocate and a main point of contact for police and other agencies, and by developing safety plans and options. The use of advocates such as IDVAs is known to have a positive impact on victim cooperation (Camacho and Alarid, 2008) as well as a moderately positive impact on quality of life, but mixed effects on recidivism in relation to physical and sexual abuse (Rivas et al., 2015).

9.5.6 Multi-agency meetings

For cases designated as high risk, it is common practice for the victim to be discussed at a meeting of agencies concerned with the issue of domestic abuse (police, probation, children's services, housing agencies, charities, education agencies, etc). In England and Wales, these meetings are known as Multi-Agency Risk Assessment Conferences (MARACs). There are more than 270 such meetings, dealing with almost 100,000 cases every three months (College of Policing, 2018; SafeLives, 2018). The intention of MARACs is to reduce the risk to victims and their children by facilitating information exchange between agencies and building plans of action. Two thirds of referrals to MARACs originate from police forces, and just over a quarter of cases are discussed repeatedly (SafeLives, 2018).

MARACs are a form of 'co-ordinated community response' (CCR), which have been explored in domestic abuse research in the UK and the US, although mostly in respect of processes (Klevens, Baker and Shelley, 2008). The original evaluation of MARACs found that 40% of cases had no further police call-outs in the 12 months after their MARAC contact (Robinson and Tregida, 2005), but other than this study, the evidence base features little robust proof of the notion that MARACs or CCRs more generally fulfil their purpose. Klevens et al. (2008) compared domestic abuse rates between 10 CCR areas and 10 areas

without CCRs and found no significant difference. Two quasi-experimental studies replicating Klevens et al.'s research reached the same conclusion (Post et al., 2008; Visher et al., 2008). Research conducted for the UK Home Office in 2011 concluded that evidence on the impact of MARACs on outcomes was quite weak (Steel, Blakeborough and Nicholas, 2011), and a number of other studies in the country have echoed that concern (McGlaughlin et al., 2014; Berry, Stanley, Radford, McCarry and Larkins, 2014). More recent experimental evidence has found that multi-agency perpetrator management approaches may offer some reduction in crime harm over a two-year follow up period (see Goosey, Sherman and Neyroud, 2017). However, this single study included an imbalance in treatment intensity which limited the precision of the findings in respect of identifying which aspect of the multi-agency approach caused the effect.

9.5.7 Perpetrator management

In 2016, the police inspectorate explicitly called for police to detail what perpetrator programmes they were operating with reference to research published by the College of Policing (HMICFRS, 2017). That research, however, found no conclusive evidence about any form of perpetrator scheme, primarily due to a lack of well-designed evaluations (Vigurs et al., 2016). The only guidance available to forces in respect of perpetrator management relates to serial and repeat perpetrators, for whom the College of Policing advises that each force should have a system for active management and monitoring which makes use of existing schemes such as Integrated Offender Management and Multi-Agency Public Protection Arrangements (College of Policing, 2018). In response, at the time of writing, many police forces were trialling perpetrator management programmes such as Drive (www.driveproject.org.uk) and multi-agency tasking and co-coordination (Davies and Biddle, 2017).

9.6 Summary

These descriptions of current domestic abuse practices are by no means exhaustive, but they cover the main components of the policing response. One might also write about legislative changes such as the domestic violence disclosure scheme, or specialist justice provisions such as specialist domestic violence courts, but they are less relevant to establishing context for this research which focuses predominantly on issues pertinent to triage and prevention.

What is starkly apparent from these descriptions is the paucity of robust underpinning evidence. Instead, it is difficult not to draw the conclusion that domestic abuse strategy is

established on a bedrock of professional judgement rather than evidence-based practices, which is arguably paradoxical given the stated ambitions of the police service. There are a number of influencing factors here. Firstly, there is little strong evidence about domestic abuse activities. The deepest evidence base is in respect of arrest, yet even that is mixed. Other cornerstone elements of the domestic abuse response (risk assessment, advocates, MARACs) have been subject to so few high-quality studies that there is no strong evidence to speak of at all. In the face of such a void, it is perhaps not so surprising that agencies have used professional experience, and in some cases low-quality studies, on which to design their responses.

The problem of domestic abuse is pressing, highly prevalent, costly and harmful. In the last two decades, scrutiny of the response to domestic abuse has grown ever sharper, culminating in a damning report of the policing response in 2014 (HMICFRS, 2014a). This engendered a prevailing imperative to 'do something' – yet little robust evidence was available to inform practice. For many domestic abuse response initiatives, once they have been piloted, roll-out seems inevitable; and once rolled out, it is unthinkable that they would ever be stopped, even to allow for control-group based trials. The domestic abuse response community has been largely reluctant to build 'denial-of-service' control groups into any forms of evaluation for fear of harming victims by doing so, and this has led circumstances in which un-evidenced responses have been implemented to the masses with no plans to rigorously test their impact on outcomes. In the few examples where this prevailing view has been overcome (see Strang et al., 2017), the results have been slow to gain traction.

Yet this status quo is not sustainable in the face of ever-increasing demand on domestic abuse services. The number of domestic abuse cases reported to police forces has increased substantially in each of the last three years (ONS, 2016a, 2017, 2018; HMICFRS, 2019), and the number of cases seen by MARAC agencies has followed a similar trend (SafeLives, 2018). These trends are unlikely to abate in the face of ongoing scrutiny from the police inspectorate and the ongoing advocacy of large national charities that conduct and publish their own research. It is equally improbable that the agencies responding to domestic abuse, especially the police, will have more resources to deal with the problem in the near future. In addition to the £1.6 bn cut from the policing budget between 2010 and 2017, and a further planned £700 million to be cut by 2021, the police service was given a £420 million bill by the UK Treasury for pensions shortfalls (Dodd, 2018), for which only temporary central funds were supplied. This situation mirrors that faced by other government agencies:

local authorities have approximately 26% less funds than in 2010 (Hulme, 2017), which affects adult social care and children's services capacity; the amount of education spending per pupil fell by 8% between 2010 and 2018 (Coughlan, 2018); and by 2020, the Probation Service budget will have been reduced by 40% over the course of a decade. All of these cuts have potential adverse implications for domestic abuse prevention.

The continuing effects of austerity do not make the development and refinement of an evidence base for domestic abuse response less relevant; they make it more so. There is a real and pressing need for agencies to better target their scarcer resources in order to achieve their desired outcomes of protecting victims. While it may be that this requires a change in attitude towards control-group-based evaluations, these need not be the only source of evidence upon which strategies are refined. Targeting evidence may be just as, if not more, useful to agencies in the current context, and this is where police data may have a crucial role to play.

10 Measuring Domestic Abuse

10.1 Chapter roadmap

The previous chapter described the environment in which domestic abuse practitioners and researchers operate. One theme that runs throughout that environment is how episodes of domestic abuse are measured. This chapter examines the two principal methods of recording domestic abuse – public surveys and police records. As the latter is the source of data for the analysis that follows, this chapter sets out the key information readers need to understand to contextualise the methodology, its limitations, and the analyses.

10.2 Victim surveys

The Crime Survey of England and Wales is commonly referred to as the best estimate of domestic abuse prevalence (SafeLives, 2018; Women's Aid, 2017). This survey is used by the Office for National Statistics to produce the official national estimates of domestic abuse prevalence. In respect of domestic abuse, the survey itself comprises two parts: interviews and a self-completion module. Until 2018, the interview process did not exactly match the cross-government definition of domestic abuse, particularly in respect of coercive and controlling behaviour, for which it included no questions. Thus, prevalence estimates were calculated using the self-assessment module, which has historically had a higher reporting rate than the interviews (ONS, 2018). The ONS estimated that 1.9 million adults (aged 16–59) experienced domestic abuse in the year ending March 2017, a rate of 6 in 100 adults, but given the caveats, this estimate was certainly below the true level. Future surveys will improve this situation, and it is anticipated that the questions introduced on coercive and controlling behaviour will increase gender asymmetry (Myhill, 2015), but for now the 'best evidence' of prevalence indicates a broadly stable trend in the last decade (see Figure 1).

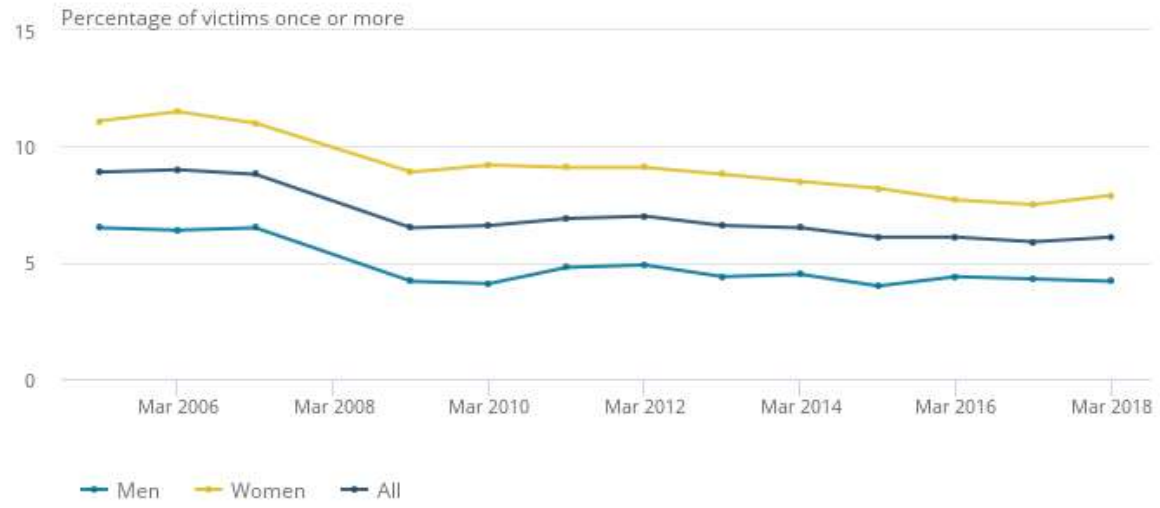

Figure 1. Prevalence of domestic abuse in the last year for adults aged 16 to 59 years, by gender[6]

The Crime Survey of England and Wales offers some points of interest to strategy-makers, for example on age or gender profile, but disaggregated data are not routinely made available to agencies to interrogate, and to date the survey has not been used to answer questions beyond the most strategic.

Whether the survey represents the 'best evidence' of prevalence is also questionable. Although the anonymous self-reporting methodology potentially overcomes underreporting challenges in the broadest sense, the notion of a large gap between official records and surveys for more serious crimes, such as violent crimes resulting in severe injuries, has been challenged (Ariel and Bland, 2019). With these views in mind, records kept by police forces are a potentially promising option for deeper exploration of domestic abuse patterns.

10.3 Police Records

As already outlined, the police service has come under considerable pressure to 'up its collective game' in respect of crime recording, and this appears to have led to a narrowing of the gap between actual and recorded prevalence. Police forces recorded 1,198,094 domestic abuse events in the 12 months ending March 2018 (ONS, 2018) – the smallest gap on record between police records and the Crime Survey of England and Wales. This is a large pool of record level data on domestic abuse, and if a larger one exists, it has not yet been discovered.

[6] Reproduced from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwales/yearendingmarch2018

Size is not everything, however, and if this data source is to be a real option in defining domestic abuse issues and calibrating responses accurately, it bears closer inspection for quality (an issue we return to later in this chapter).

The Home Office made domestic abuse recording mandatory for police forces in 2015, following the HMICFRS' thematic inspection of domestic abuse, which called for an improvement in data recording and the establishment of a national domestic abuse dataset (HMICFRS, 2014a). Despite this, most police agencies in England and Wales had already been recording domestic abuse under the national definition as part of their own internal processes. Domestic abuse has been recognised as a core area for policing in England and Wales for more than a decade, and most forces had already established specialist units and specific processes to tackle this crime type. The Domestic Abuse, Stalking and Honour-Based Violence (DASH) risk assessment process meant that forces could already distinguish domestic cases to some extent, albeit without audit or regulation; no agency had set or inspected police data standards in respect of domestic abuse classification until the Home Office established such standards in 2015.

Police domestic abuse data are perhaps the most obvious choice for the development of evidence for the purpose of targeting resources, yet these have no real reputation to speak of as a source for empirical research. Even though the police service introduced a national standard for crime recording in 2002, several governmental reports published in subsequent years were critical of the quality of police–recorded crime data (PASC, 2014), leading the House of Commons Public Administration Select Committee to conclude in 2014 that they were no longer a reliable source. Shortly thereafter, the UK Statistics Authority withdrew the designation of police–recorded crime statistics as a 'National Statistic'. This led the police inspectorate to conduct a nationwide series of inspections of crime data quality, on the basis of which it overwhelmingly concluded that the police service in England and Wales was substantially under-recording offences, particularly those of an interpersonal nature (HMICFRS, 2014b). This inspection contained a specific recommendation that police forces submit domestic abuse records to the Office for National Statistics to use in the creation of a national dataset aimed in particular at addressing the issue of defining repeat victimisation. This national dataset was submitted from the beginning of the 2015/16 financial year, and the overall result of the inspection programme has been a sharp rise in overall levels of recorded interpersonal crimes.

Critics of police data in domestic abuse research commonly have three main complaints (Brimicombe, Brimicombe and Li, 2007; Brimicombe, 2018; PASC, 2014). Firstly, they argue that the data are difficult to assimilate across force boundaries, being held on numerous different systems and in numerous formats. Secondly, the data are often 'incomplete', which is a particular problem for domestic abuse researchers because collation relies on a process known as 'flagging', whereby officers mark offences as domestic once they have identified them as meeting the definition. Thirdly, and most importantly, critics point out that police–recorded domestic abuse represents only a small proportion of actual crime, rendering it inappropriate as a source from which to draw conclusions about the crime type as a whole. All of these are relevant and valid concerns, but each can be overcome.

The advent of the national dataset has meant that police forces are required to keep certain fields of data, and modern data-cleaning techniques mean that more data matching is now possible than ever before. Forces have improved the consistency of their record 'flagging', under scrutiny from the inspectorate and in anticipation of future inspections. Among the 23 forces subjected to data quality inspections in 2017 and 2018, the mean accuracy of domestic abuse records exceeded 80%.[7] With domestic abuse now designated as a statutory data return, the national network of crime registrars and their audit work come into play, further increasing the checks and balances on these data.

Finally, the most important criticism: that most domestic abuse is not reported to the police. It is difficult to find a precise measure on this, but it is commonly accepted among domestic abuse researchers that this is the case (Brimicombe, 2018). The only practicable way to check the level of underreporting is to compare the Crime Survey of England and Wales (CSEW) prevalence estimates compiled by the Office for National Statistics with the level of police-recorded crime. The national dataset now provides this comparison, as shown in Table 1.

[7] Figures retrieved from https://www.justiceinspectorates.gov.uk/hmicfrs individual reports into police force crime recording inspections.

Table 1. Comparison between CSEW and police–recorded domestic abuse, 2015 to 2018

	2015/16	2016/17	2017/18
a. Number of victims estimated by CSEW	2,000,000	1,913,000	2,000,000
b. Number of crimes and incidents recorded by police	1,031,120	1,068,200	1,198,094
Ratio a:b	1.94:1	1.79:1	1.67:1

It must be stressed that this is not a like-for-like comparison; the CSEW estimates the number of victims in a twelve-month period, regardless of how many crimes they have experienced. Conversely, the police-recorded figure reflects the number of events reported to them in a twelve-month period, in which victims may appear more than once. Police records also include victims over the age of 59, which is the maximum age considered by the CSEW. These issues aside, the gap in the ratio of CSEW-estimated victims to police-recorded crimes and incidents is clearly closing. While this might be explained by improving trust in the police on the part of victims once contact has been established, the fact remains that the police data pool is not insignificant by any means. Were it considered to be a sample of the overall domestic abuse population, the general confidence interval would likely be low, and while it is clearly not a sample in this sense (as it is not randomly drawn or stratified in any way), this notion is illustrative of the potential overall power of this dataset. Add to this the fact that the common rulebook and audit infrastructure make police records a common language understood and translatable anywhere in the country, and these data offer not just the best opportunity for quantitative criminologists to explore domestic abuse, but also the most practicable opportunity by far. In a resource-restricted environment, there is compelling reason to examine this ready-made data resource.

10.4 What problems to measure?

Thus far we have established that domestic abuse is a major concern for public and charitable agencies in England and Wales, and that though their current response is extensive, it is neither based on rigorous evidence nor is the response rapidly accumulating that evidence. Yet with scrutiny undiminishing and the volume of work still increasing, while funding for resources continues to reduce, the need for evidence is as acute as ever. With agencies

seemingly reluctant, on ethical grounds, to engage in widespread testing of the kind that would provide clear outcome comparisons between different tactics, targeting evidence – identifying who, where and what to address – is possibly the most pressing kind of information required. If decision-makers in domestic abuse agencies cannot continue to provide an equal service to all, to whom should they provide a disproportionate service, and hope to secure the best outcomes? This is the prominent contextual challenge that frames this research, and we have argued that police data, as the richest single source of information on domestic abuse victims and perpetrators, hold the best potential to provide such evidence.

First and foremost, however, we need to understand what outcome is sought by practicing agencies. This perhaps appears a simpler challenge than it really is because there is no single national domestic abuse strategy, and as such, it is worth briefly examining the stated strategic outcome aims of some of the key organisational stakeholders.[8]

Table 2. Stated aims of selected national domestic abuse strategies

Agency/strategy	Stated aims
UK Government Violence Against Women and Girls Strategy 2016–2020	Continued decreases in the prevalence of domestic violence More victims helped into long-term independence
SafeLives	To end domestic abuse for good
Women's Aid	Safety, freedom and independence
Ministry of Defence	Reduced prevalence and impact of domestic abuse and increase safety and wellbeing of all those affected

Modest though this selection is, it is indicative of the multitude of local strategies found on local authority, police and crime commissioner websites. There is a consistent common sentiment throughout – to make victims safer – but a striking lack of agreement on specific and measurable outcomes, ranging from the comparatively conservative 'decrease in prevalence' to the total and permanent cessation of domestic abuse. Our interest here is not to

[8] Note the focus on outcome aims. Many strategies include output aims in their content, for instance, increase service availability', which is a means to an end (output) rather than the end itself (outcome). Table 2 focuses only on those aims seen as relating directly to outcomes.

evaluate the merits of these strategies but rather to determine how one might develop a system of measurement, on the presumption that measurement is a prerequisite for systematic targeting, testing and tracking. The obvious candidate is prevalence, which is implied or explicitly stated in most domestic abuse strategies. Easy to understand though it may be, prevalence requires survey estimates and discounts the differential nature of crimes. Prevalence could be decreased by 50%, but if the remaining crimes included an increase in homicide, rape and serious assault, it would be illogical to argue that the outcome was a successful one.

Consequently, to throw light on differential patterns of harm, we need to be able to define and measure the concept of harm. Furthermore, the instrument we use for this purpose must be complementary to the source of our data (police records). The next chapter focuses on these issues.

10.5 Summary

Domestic abuse is measured in two main ways – public surveys and police records. It is commonly accepted that the latter source does not represent all domestic abuse, though recent trends suggest the gap between actual and recorded crimes is falling. Nonetheless, police records are the only large data source which identify individuals and thus support analytical procedures that may provide targeting insights. The largest survey in England and Wales yields data which are aggregated and anonymous and therefore cannot be used in the same way. Police datasets are also very large and cover categories of victims not reached by the official crime survey. However, whichever tool were to be used, there is still a clear and present need to identify an instrument capable of differentiating between levels of harm. Most domestic abuse strategies are focussed on harm reduction in one form or another, yet there is no established mechanism for tracking this notion.

11 Measuring Harm

11.1 Chapter roadmap

This chapter considers the definition of harm within the context of responding to domestic abuse crimes. Specifically, it appraises a selection of instruments which could potentially be used to analyse patterns of harm in police domestic abuse datasets. To this end, the chapter begins with a brief discussion of the general concept of harm and its measurement, then reviews a selection of instruments that have been developed in recent years. The chapter concludes by applying three tests, developed by Lawrence Sherman and Peter and Eleanor Neyroud (2016), to form a recommendation about which instrument to use in the subsequent analysis.

11.2 What is harm and how is it measured?

Meaningful analysis of domestic abuse aimed at improving targeting strategies must surely differentiate between crimes based on their relative 'harm'. In criminological context 'harm' is traditionally described as an emotional, psychological, financial, societal or physical impact (see Adler, 2001 or Sparrow, 2008 for examples of discussions about the definition of harm in the context of crime). Accordingly, a number of harm measurement frameworks and tools have emerged in the last three decades. The underlying premise for these instruments is this: not all crimes are the same and treating them as such skews analysis and interpretation of policing issues, leading to the misallocation of resources (Sherman, 2007, 2013; Sherman, Neyroud and Neyroud, 2016).

Any promise that police data holds might be undermined in the absence of a harm measurement instrument or with the selection of an inappropriate one. The need for such a tool in the analysis of domestic abuse has been hinted at in earlier chapters. Not all domestic abuse is the same; within the high volume are smaller numbers of severe and serious cases (Bland and Ariel, 2015; ONS, 2017, 2018), yet as explored in Chapter 8, the police response, marked by reduced capacity, invests relatively high resources in all cases. Analysing only aggregated trends in police data will offer at best only a limited remedy to this problem, whereas viewing the data through a lens which differentiates by harm could be the key to answering many research questions of substantial practical value. If we can filter the most harmful cases, we stand a better chance of understanding them, designing treatments for them, and possibly even forecasting them before they become harmful. Such a cohort may

also offer the best potential for detecting effects in future domestic abuse experiments (see Sherman, 2007, regarding the promise of 'the power few').

For over two decades, criminologists have debated the concept of the measurement of harm, and its potential role at the heart of anti-crime strategies. In recent years the debate has shifted into the professional arena and is now characterised by policing strategies targeting 'vulnerability' (College of Policing, 2016; HMICFRS, 2015) as well as calls from senior government officials for the adoption of 'public health models' to deal with specific issues such as knife crime (see *The Independent*, 2018). At the core of these debates is the definition of 'harm' and its practical dependencies, such as how it is measured and translated into practice. These are complex issues that do not currently have wholly satisfactory answers, with the debate around the concept of harm ranging across a wide spectrum of views. At one extreme, some critical criminologists challenge the very nature of our understanding of crime, claiming that it has no substantive ontology (Hillyard and Tombs, 2007). This argument centres on the proposal that 'harm' is a more appropriate target for service delivery and research than the restrictive notion of crime. It suggests that the majority of police-recorded crime is 'petty' and inconsequential in terms of harm, while many of the more serious harms in society are omitted (Box, 1983). This end of the harm debate spectrum is highly theoretical, to the point of impracticality, making it so obscure to practitioners as to be virtually invisible. This is not to say that it is invalid, but in terms of identifying a lens of harm with which we may realise the potential power of police data, it is hardly helpful.

However, this is not the full extent of scholarly discourse on harm and its potential influence on law enforcement activities. Sparrow (2008) specifically focuses on the measurement of harm at a practical level, drawing on invented scenarios to illustrate his points. One of Sparrow's central recommendations is to 'pick important problems and deal with them' (Sparrow, 2008, p. 5), which is a variation on the theme of Sherman's evidence-based targeting (Sherman, 2013). Sparrow notes that successful practitioners take time to understand where risk (as a construct of potential harm) is concentrated. The missing link to which Sparrow dedicates a considerable number of pages, is the role of analysis in the quantification and subsequent definition of harm problems. Drawing on Goldstein (1990), Sparrow cogently argues for a focus on the 'middle-layer' of issues – that which lies between the response to individual incidents and thematic strategies. This is precisely the domain of this research and explains why an entire chapter is devoted to the identification of the right instrument with which to measure harm.

This chapter chronologically details the development of harm-measurement tools in the last three decades, paying specific attention to the methodology of their development. In this respect, the tests summarised in Sherman, Neyroud and Neyroud (2016) are used for the purposes of benchmarking. These tests are designed on the basis that, to be practically effective, any harm-measurement instrument must satisfy the following criteria:

1) The democracy test: Does the instrument satisfy conflicting arguments via democratic means?
2) The reliability test: Can the measure be consistently applied to different units of analysis, remaining consistent over time?
3) The cost test: Is the instrument available at no or low cost?

In this chapter, the third test is used as a proxy measure for 'practicability' by assuming that, if the tool is available at no or low cost, then it must be easy to implement. Our concern here is not for the difficulty of this research, but for its potential replicability among practitioners. Bland and Ariel (2015) utilised the Cambridge Crime Harm Index (CCHI), for which these three tests were conceived (see Sherman, Neyroud and Neyroud, 2016), but for which no thorough selection process was undertaken. Since then, other tools have emerged to complement or compete with the CCHI and the tools which preceded it, and so here we will critically and objectively evaluate each of the possible options and draw a conclusion regarding which harm-measurement instrument to proceed with in the subsequent analysis.

11.3 Review of harm measurement tools

Before considering proposals for specific instruments, it is worthwhile briefly reviewing the history of the development of such tools so as to place them in context and better refine the method of selection. Four broad categories of 'crime harm' measurement tool have emerged in criminological research: (1) public perception–based indices; (2) cost of crime indices; (3) sentencing-weighted indices; and (4) theoretical constructs. In this section, we will review the development of each, beginning with the seminal work of Sellin and Wolfgang (1964).

At the outset of this brief history, emphasis must be placed on the fact that there is no consistent definition of 'harm' shared by these tools because none has been agreed on in previous research. Many of the tools examined purport to measure 'severity' or 'crime seriousness', and for the purpose of this research we treat these as proxy terms for harm on the following logic: if crime X is more serious or severe than crime Y, then it follows that X

is more harmful. This logic necessitates the precondition that the tool uses the concept of 'harm' as a key influencing factor in the determination of seriousness, and specific attention is paid to this in the evaluation of the specific tools.

11.3.1 Public perception–based tools

In 1931, Thorsten Sellin identified that aggregating simple counts of crime was a poor measure of criminality:

> Criminal statistics have not yet reached a uniformly high stage of development, however, and this in part accounts for the frequency with which they are abused. (Sellin, 1931, p. 10)

It was not for another 33 years, however, that Sellin and his colleague Marvin Wolfgang would develop the first major instrument to differentiate between the seriousness of different forms of crime by an empirical mechanism and present it as a supplement to simple aggregated counts (Sellin and Wolfgang, 1964). Their method was to sample students, judges and police officers, asking them to rate the severity of 141 different criminal scenarios which mirrored the Federal Bureau of Investigation's Uniform Crime Reports, the prevailing measure of crime seriousness in the USA at the time. Sellin and Wolfgang's framework comprised two major classes of crime, subdivided into subclasses (see Table 3).

Table 3. Sellin and Wolfgang's (1964) severity typology

Class I		Class II	
A	Bodily injury	D	Intimidation with threat of violence
B	Property theft	E	Intimidation with threat of damage
C	Property damage	F	Primary victimisation (to a person)
		G	Secondary victimisation (to a business or organisation)
		H	Tertiary victimisation (e.g., to the state or community)
		I	Mutual victimisation (e.g., adultery, other consensual illegal acts)
		J	No victimisation (juvenile offences)

For each of their 141 scenarios classified within these codes, Sellin and Wolfgang developed vignettes. These were presented to their participants, who were asked to rate them. These vignettes were the source of much of the early criticism of Sellin and Wolfgang's work. Rose (1966) highlighted that the vignettes were often inconsistently presented, and the individual characteristics of victims or perpetrators were vital to the perception of seriousness because of the role of public stereotyping. This early critique drove right to the heart of a crucial matter in severity measurement; the issue of subjectivity in the perception of seriousness. This issue has been the subject of much scholarly debate throughout the years since Sellin and Wolfgang's work and agreement remains elusive (see Cohen, 1988; O'Connell and Whelan, 1996; Rossi, Simpson and Miller, 1985; Styliannou, 2003 and Sherman, Neyroud and Neyroud, 2016 for examples of this debate). In Wolfgang et al.'s (1985) follow-up work, in which the National Survey for Crime Seriousness was developed, the authors rejected Rose's criticism. However, their rejection appears somewhat selective, being based on a study of 206 students comparing six scenarios which concluded that information about intent and culpability formed a critical component of a person's judgement of severity (Riedel, 1975). Reliance on a single small-sample experiment for such an important finding is highly questionable, and indeed the study's conclusion was questioned by Sebba (1984) and later contradicted in a larger experiment conducted in Israel (Fishman, Kraus and Cohen, 1986).

Nonetheless, Sellin and Wolfgang's work in 1964, and later Wolfgang, Figlio, Tracy and Singer's work in 1985, became the benchmarks for the measurement of crime severity from the 1970s to the 1990s, spawning a number of replications and spin-offs (Akman and Normandeau, 1968; Blumstein, 1974; Epperlein and Nienstedt, 1989; Fleming, 1981; Lynch and Danner, 1983; Parton, Hansel and Stratton, 1991; Rossi et al., 1974). While the findings and theoretical debates may have varied in this body of work, a consistent set of characteristics have emerged which demarks this strand of harm measurement tools from the other three and continues to pervade their overall debate. The public perception–based strand of instruments is characterised by three main aspects, all of which are relevant to the potential selection of such a tool for this research: (1) questionnaire design, (2) levels of measurement and (3) 'additivity'. Styliannou (2003) gives an excellent full description of these, but it is worth summarising the main points here before evaluating the potential use of this type of tool for the purposes of this research.

Firstly, the issue of questionnaire design is fundamental to perception-based instruments. Much prior research has found similarities in the ways in which different groups of individuals conceptualise seriousness (McCleary et al., 1981; Pontell, Granite, Keenan and Geis, 1985; Rossi et al., 1974). Warr (1989) distilled these factors into a simpler equation: seriousness as a product of perceived harmfulness and perceived wrongfulness. Other researchers already disagreed with this on a fundamental conceptual level (Blum-West, 1985; Hansel, 1987), and the matter at hand here can be described as follows: if scholars cannot agree on a definition of seriousness, and further still can demonstrate that external factors such as stereotypes may colour perceptions, then how definitive and representative can the output of severity surveys ever be? This is compounded somewhat by the fact that, by definition, this type of measurement instrument draws on probability samples, and in most of the literature to date, most of those samples consist of university students. Therefore, the selection of a tool from the public perception–based strand must carefully consider the structure of the questionnaire on which the output is based, specifically with regard to:

1) the typology of, and balance between the scenarios presented,
2) the constitution of the population from which the sample is drawn, and
3) the order and presentation of the questions.

The second important consideration is the method by which the instrument measures severity. Researchers working with public-perception tools have three primary methods: (1) ordinal/categorical scales, (2) magnitude estimation scales and (3) matched-pairs comparisons. Wolfgang and Sellin (1964) favoured magnitude estimation; the researchers invited participants to judge severity by comparing each scenario to a benchmark scenario (e.g., twice as serious, one hundred times as serious), arguing that this better reflected the construct of the participant panel, and not that of the person designing the measurement method. Their thoughts on this subject have been echoed many times (Bridges and Lisagor, 1975; Evans and Scott, 1984; Figlio, 1975; Rossi and Henry, 1980; Wolfgang et al., 1985), yet the method has been challenged by Miethe (1991) and Parton et al. (1991), who cautioned that it would be necessary to train participants in order to ensure the reliability of the measure. Fishman et al. (1986) also remarked on the training and level of competency required by panel members. In the consideration of the appropriate model for this research, the measurement scale is critical. The relative magnitude of harm is an essential component of such a tool if this research is to properly expose trends in harm concentration, escalation

and more. This dilemma is best illustrated by means of a simple example. Is a 'common assault' domestic crime, wherein the victim is assaulted but sustains no injury, less serious than a 'grievous bodily harm' crime in which the victim suffers a serious physical injury? The latter is more serious in this author's estimation – but by how much? Is it twice as serious? Ten times as serious? A hundred times as serious? This detail matters in practice: how many common assault crimes constitute the harm of one grievous bodily harm crime?

This leads us to the third important consideration among perception-based tools, which Stylianou (2003) labels 'additivity'. Sellin and Wolfgang's original premise (1964) was that two crimes of the same kind, committed either repeatedly or at the same time, were empirically equivalent to separate instances of the crime (e.g., committed by different people). Though this view was supported by Wellford and Wiatrowski (1975), it was challenged to varying degrees by Pease, Ireson and Thorpe (1974), Wagner and Pease (1978) and Gottfredson, Young and Laufer (1980) on the basis of interactive considerations on the part of panel participants. Ignatans and Pease (2015), in developing a proposal for a UK harm index based on victims' perceptions of seriousness as reported in the Crime Survey of England and Wales, associated 'additivity' with people's judgements of seriousness, harm, and culpability. They argued that, in surveys which sample both single and chronic victims, the relative seriousness is factored into the output weighting. This is another key test in selecting the right tool for this research; while individual classification of harm is desirable (taking into account each individual circumstance of the victim and offender), it is unlikely to be practicable. Therefore, an assumption of broad 'additivity' is important to keep in mind when selecting the tool to ensure consistency of measurement and practicability (two of the three key tests set out by Sherman, Neyroud and Neyroud, 2016).

11.3.2 Economic harm–based tools

A second strand of harm metric tools is composed of those which classify harm according to economic harm or financial cost. While cost is one aspect of severity considered by many of the public perception–based tools discussed in the previous subsection, this strand can be distinguished by the use of currency values as the output harm metric; these tools literally put dollars and cents or pounds and pence forward as the denominator of harm.

The depth of previous research covering this strand of tools is somewhat lighter than for perception-based tools. The most relevant models of interest to this research took their cues from a modest range of research from the last century. Cohen (1988) first introduced the

broad concept of supplementing 'costs incurred' estimates with information about 'pain, suffering, and fear caused by crime' (Cohen, 1988, p. 1). Cohen, whose research focused on the United States, observed that, up to the point of his publication, most efforts to measure the cost of crime had focused on actual financial costs. In this respect, earlier efforts cannot be considered as even proxy measurements of harm. Subsequent work involving Cohen concluded that adding the costs of pain, suffering, and fear more than quadrupled the cost of crime (Miller, Cohen and Wiersema, 1996), with the increase being attributed primarily to violent crime. Subsequently, other cost models were developed in France (Palle and Godefroy, 1996) and Australia (Walker, 1997).

In England and Wales, cost-of-crime models took a leap forward with practitioners after the Home Office published a research paper detailing a thorough costing model to be used in performance management (Brand and Price, 2000). The Home Office's tool estimated costs related to the anticipation of, response to, and consequences of crime, within which values were assigned for emotional and physical impact. The tool was configured around the framework of notifiable offences in England and Wales at the time, grouping together categories into eight classifications. The cost estimates themselves were the product of a complex combination of victims' surveys, commercial surveys and industrial estimates. By the authors' own admission, the methodology for developing the costs of emotional and physical impacts required improvement. The Brand and Price methodology used public perception of the costs a person would be willing to incur to avoid a road traffic accident as a proxy for the impact of crime. Clearly unsuitable, this specific aspect was addressed in a subsequent Home Office paper (Dubourg, Hamed and Thorn, 2005) which assessed a range of surveys to identify the prevalence and severity of health conditions emanating from crimes, then transposed these to estimated losses of quality-adjusted life years. This study gained some traction among researchers and professionals in subsequent years (Ignatans and Pease, 2015; Welsh, Farrington and Gowar, 2015). Despite this, the tool was not updated for more than a decade, becoming obsolete due to the absence of official inflation adjustment. The use of cost-based models has been criticised by some as having low practical utility for practitioners because of the difficulty inherent in assessing a meaningful monetary value for many crimes as well as the need to constantly adjust for inflation (Ratcliffe, 2016). The Home Office refreshed the model once again in 2018 (Heeks et al., 2018), concentrating primarily on victim-based crimes.

11.3.3 Sentence-based tools

Sentence-based tools weight crimes on the basis of their respective punishments. In this sense, there are two main types of sentence-based tool: those which take their weightings from sentencing guidelines and those which take them from actual sentences imposed. These types of tool are a relatively recent development among criminologists, but they have become popular among analysts and researchers wishing to assess harm (see Barnham et al., 2017; Bland and Ariel, 2015; Dudfield, 2016; Sherman, Bland, House and Strang, 2017). Recent developments can be traced back to Sherman's call (2007) for a mechanism for weighting crimes in order to target experiments at what he called 'the power few' – units (whether people or places) to which were attributed the greatest harm. Sherman's case was stated on the basis that such a cohort may offer the best opportunity for experimental criminology to detect effects in treatments.

Two years later, the first such model emerged with the composition of the Canadian Crime Severity Index (Wallace et al., 2009), though it should be pointed out that the catalyst for its development was not Sherman's 2007 paper but rather a 2004 call from the Police Information and Statistics Committee of the Canadian Association of Chiefs of Police (CACP), which requested a new method of reporting crime statistics from Statistics Canada, the Canadian equivalent to the UK Office for National Statistics. Their intention in seeking such a tool was different to Sherman's; they wished to be able to detect changes in crime rates in a more nuanced way than mere aggregated crime statistics could portray, but this difference is somewhat semantic – both Sherman and the CACP sought variations on Sellin's original premise: not all crimes are equivalent. In the Canadian Crime Severity Index, crimes recorded by the police are assigned weightings based on the mean sentences given in Canadian courts over the preceding five years (Babyak et al., 2009; Babyak et al., 2013). Almost ten years on, the tool has become a mainstream national statistic (see https://www.statcan.gc.ca/eng/sc/video/csi), and annual analyses of Canadian crime rates are presented using it.

Sherman restated his argument for a 'crime harm index' repeatedly after his original 2007 paper (Sherman, 2010, 2011, 2013) and in this time developed processes for what would later become the Cambridge Crime Harm Index (Sherman, Neyroud and Neyroud, 2016). Like the Canadian index, the Cambridge Crime Harm Index is aligned to police-recorded crime classifications, but instead of taking individual weightings from average sentences, it takes them from the minimum sentences set out in guidelines given to judges

and magistrates in England and Wales. Such a method is enabled by the fact that such guidelines exist at all, which they do not in every country.

The development of the Cambridge index was the catalyst for a relative explosion in sentence-based crime indices. The 'Sentencing Gravity Score' (Ratcliffe, 2015) proposed a similar method of taking a cue from guidelines to establish weight, but instead of using the number of days in prison as the unit of output, as do the Cambridge and Canadian indices, this model used a 14-point scale of severity, derived from scores assigned by the Pennsylvania Commission on Sentencing, which was adopted in 1997. Ratcliffe favoured this method due to its specificity and independence from police input (Ratcliffe, 2015). The Sentencing Gravity Score broadly correlates with homicide rates; where the index score was high in Pennsylvania, it followed that the homicide rate was high, although this relationship deteriorated at lower-level geographic units once traffic accidents and other proactive measures were included. In this respect, Ratcliffe's model departs from Sherman, Neyroud and Neyroud's in that the latter authors argued for the removal from the index of crimes recorded as a result of proactive police activity (e.g., drug possession crimes). While Ratcliffe wished to build a model reflective of wider police activity, Sherman et al. were primarily concerned with the differential effects such inputs may have, being largely dependent on the individual proactive capacity or working practices of individual agencies. Both points have merit; the Sentencing Gravity Score was developed specifically for use in one state and tailored to maximise the chances of operationalisation, while the Cambridge Crime Harm Index sought to establish a measure which could be meaningfully transposed across police force boundaries.

Other models influenced by the development of Sherman et al.'s model more closely mirrored it in respect of output (i.e., scores reflective of the number of days in prison). Notably, replications of the Cambridge Crime Harm Index have evolved in Denmark (Andersen and Mueller-Johnson, 2018), Sweden (Rinaldo, 2015), California, USA (Mitchell, 2016), Australia (House and Neyroud, 2018) and New Zealand (Curtis-Ham and Walton, 2018), Each of these studies responded to the challenge of a lack of standard sentencing guidelines in their countries of focus in different ways. In New Zealand, average actual sentences were used and applied to all crime types, including those which were the product of proactive policing activities, to enable users of the index to choose those crime classifications which best suited their needs. In Australia, researchers evaluated the possibility of using maximum sentences before rejecting the method because of reduced variability (Kwan, 2016,

as cited in House and Neyroud, 2018). House and Neyroud attempted to survey the judiciary, but in light of a low response rate, opted to use a variation on the average of actual sentences, considering the average sentence of first-time offenders only. This method resembles something of a hybrid between the Canadian and Cambridge models, using real sentencing data but only in cases where lower sentence tariffs are normally applied. In Sweden, researchers had more success in surveying judges (Rinaldo, 2015), but in Denmark this method was rejected owing to judges' lack of specialisation in criminal law. Instead, Andersen and Mueller-Johnson (2018) surveyed Danish prosecutors, asking them to rate 43 crime types and controlling for inter-rater reliability.

The most significant development in sentence indices emanating from the Cambridge Crime Harm Index is the publication of the Crime Severity Score by the UK Office for National Statistics (ONS, 2016b). Published initially as an 'experimental statistic', the tool was intended by the ONS to complement, rather than replace, aggregated statistics. The Crime Severity Score draws directly on the Cambridge Crime Harm Index in spirit but employs average actual sentences (for a five-year period; December 2011–December 2015) over minimum guideline sentences. In seeking an objective measure, the ONS disregarded sentencing guidelines owing to too many omissions in the full range of crimes recorded by police. The ONS encountered some methodological challenges in certain crimes with low sample sizes, even in a five-year timeframe, and may extend the timeframe in future to address this issue (ONS, 2016b). The primary output difference to the Cambridge Crime Harm Index is that most offence types are shown to be more serious with the Crime Severity Score (see Ashby, 2018) because of the influence of aggravating factors. The calculations for 'days in prison' equivalency of community sentences and fines are very similar, though they differ slightly in terms of calculation specifics. The ONS intends to update the index every five years to reflect changes in sentencing values. The publication of the Crime Severity Score brought crime indices to the attention of the mainstream media in the UK for the first time (Shaw, 2016; Evans, 2016).

11.3.4 Theoretical-framework tools

The development of theoretical models is confined to a single study. Reflecting on the general progress of harm measurement among the criminological community, Greenfield and Paoli (2013) determined that not much had been done to establish definitive, systematic measurement instruments. While their paper predates the rapid rise in the number of sentence-based indices, it is difficult to challenge Greenfield and Paoli's central premise. In

the absence of such a tool, they proposed their own framework, while recognising 'major conceptual and technical challenges' (Greenfield and Paoli, 2013, p. 865), based on the subjectivity of defining harm, which they argued was particularly difficult given the infinite nature of the subject, the legitimacy of the source of its measurement, and the extent to which the tool can be quantified and standardised. Their solution to these problems was to develop a highly complex overall model (see Figure 2).

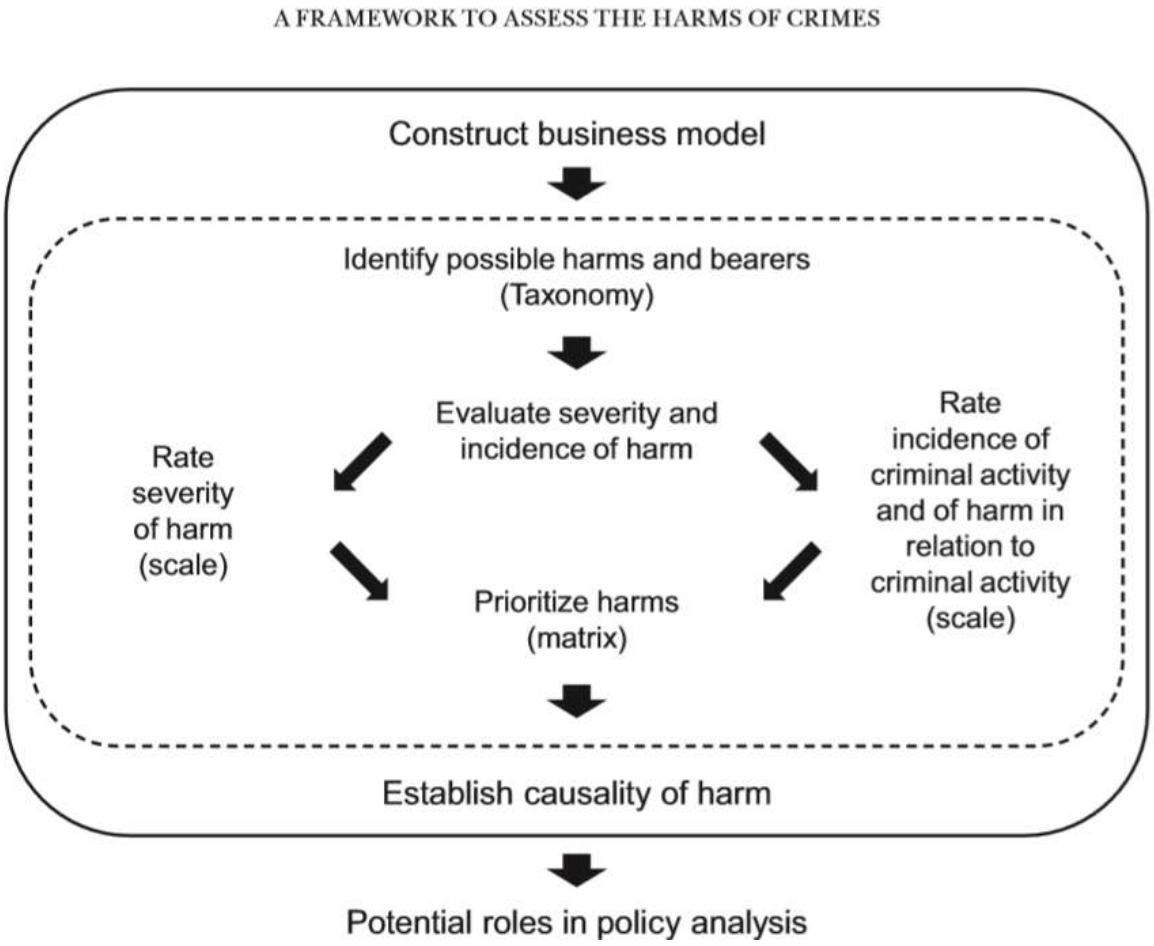

Figure 2. Greenfield and Paoli's harm assessment process, as published in Greenfield and Paoli (2013) and first published in Paoli et al. (2013)

In practice, the tool requires the identification of the bearers of harm, and the type of harm inflicted according to a taxonomy based on the work of von Hirsch and Jareborg (1991). It also requires the user to evaluate the severity and incidence of each type of resulting classification. Each of these factors has its own scale, and the positions on each of these determines the position of the type of harm on an overall prioritisation matrix. The authors advised that the determination of the positions on these scales should be made by a panel, with the output being the average of the results.

Though this model sets out a comprehensive framework, the challenges associated with its implementation are, in the authors' own words, 'daunting', with the result that it has seen little operationalisation. While it provides a recipe for a theoretically sound tool, it does not provide a prescriptive model that can be fully evaluated against the tests set out, and for this reason it is not examined further.

11.4 Assessing which tool to use

From this brief history of harm measurement tools, it is evident that we have three viable options: perception-based tools, economic-based tools and sentence-based tools. To make a thorough assessment of suitability, a viable candidate from each strand should be evaluated, and to this end each option is assessed against the following criteria: (1) Is the tool openly accessible to researchers at no cost? (2) Does the tool apply to the legal context of England and Wales? and 3) Can the tool be practically applied to police-recorded crime datasets? In applying these tests, four tools have been selected for deeper assessment, as summarised in Table 4.

Table 4. Viability assessment of harm measurement tools

Tool name	Author(s)	Open access?	Apply to England and Wales?	Apply to police datasets?
Cambridge Crime Harm Index	Sherman, Neyroud and Neyroud (2016)	☑	☑	☑
Canadian Crime Severity Index	Wallace et al. (2009)	☒	☒	☒
CSEW Victim Seriousness Judgment	Ignatans and Pease (2016)	☑	☑	☑
Home Office Economic and Social Costs of Crime	Heeks et al. (2018)	☑	☑	☑
ONS Crime Severity Score	Office for National Statistics (2016b)	☑	☑	☑
Sentencing Gravity Score	Ratcliffe (2015)	☑	☒	☑
Severity Typology	Sellin and Wolfgang (1964)	☒	☒	☒

While it is obvious that tools applying to countries other than England and Wales should ordinarily be discounted, we have included the Canadian Crime Severity Index, the Sentencing Gravity Score and the Severity Typology in this description to illustrate the points on which they fail to meet the viability criteria. It is not impossible to apply harm index outputs across national boundaries in a meaningful way (see Sherman et al., 2016). The problems with these options lie primarily elsewhere. The methodology for the Canadian Crime Severity Index is not widely available, but more importantly, being based on sentences issued by Canadian courts, it is blatantly unsuitable for the purpose of analysing domestic abuse in England and Wales. Ratcliffe's Sentencing Gravity Score is potentially more palatable as the gradings have been made public, and with fewer of them, the individual values are arguably less important, but the fact remains that it is guidance clearly not applicable to England and Wales. The Severity Typology developed by Sellin and Wolfgang is, by now, too old to have much contemporary relevance, and it too reflects views collected from a population not especially related to 21st century England and Wales.

This leaves four clear choices for closer scrutiny: the Cambridge Crime Harm Index, the Crime Survey of England and Wales Seriousness Judgement, the Home Office Economic and Social Cost of Crime model, and the ONS Crime Severity Score. In this subsection, each is analysed in further detail against the criteria set out in Sherman, Neyroud and Neyroud (2016) and reiterated in Ignatans and Pease (2016) and Curtis-Harm and Walton (2018). To assist the reader in tracing the logic of this analysis, Table 5 explains the scales against which each tool is assessed.

Table 5. Criteria and scales for assessing harm measurement tools

Grade	Resolves conflict democratically	Demonstrates reliability	Cost-effective (including practicable)
Strong	Offers a clear method for the resolution of conflicting views that is demonstrably democratic	Output measure can be applied to different units of analysis and remains consistent for >10 years	The tool can be replicated at no cost
Moderate	Offers a method for the resolution of conflicting views which is clear, yet of questionable or invalid democratic intent	The output measure can either be applied equally to different units of analysis, or remains consistent over time (<10 years)	The tool can be replicated at low cost (nominally, a one-off investment of less than £10,000)
Weak	Offers an opaque, or no method for resolution of conflicting views	The output measure is consistent only for a limited time (up to one year) and/or can be applied to only a limited set of units	The tool either cannot be replicated or can be replicated only at high cost

The following paragraphs assess, in turn, each of the four viable tools against each of these criteria, with a brief recap of the methodology of each tool.

11.4.1 Cambridge Crime Harm Index

Methodology: Weights crimes by the number of days in prison (or equivalent) each Home Office crime classification would attract under the minimum sentencing guidelines provided to magistrates and judges, excluding all aggravating and mitigating factors.

Conflicting views: Resolves conflicting views on seriousness in its core methodology. Sentencing guidelines are produced by the UK Sentencing Council, which undertakes research into policy and legal issues, drafts a guideline and consults thereon with statutory bodies, criminal justice professionals and the general public (Sentencing Council, 2018). Following synthesis of the responses, final guidelines are published. The Council itself is an independent, non-governmental body.

Reliability: Sentence tariffs can be applied to multiple units, people, places or time periods. Sentencing guidelines rarely change, and new offences such as coercive control have published guidelines.

Cost: The Cambridge Index is freely available. Being aligned to Home Office crime classification codes, it can be easily and quickly cross-referenced with police datasets.

Table 6. Suitability assessment: Cambridge Crime Harm Index

Resolves conflict democratically	Demonstrates reliability	Cost-effective (including practicable)
STRONG	STRONG	STRONG

11.4.2 Crime Survey for England and Wales victim seriousness judgement

Methodology: Weights crimes on a scale of 1 to 20 as judged by respondents to the Crime Survey for England and Wales (CSEW). In series crimes, the rating applies only to the most recent event. It encompasses all forms of crime, including those that are not reported to the police. The CSEW caps the number of rated crimes at five. It excludes business crimes, 'victimless' crimes (e.g. drug offences), and crimes against children.

Conflicting views: Not specifically addressed. The CSEW draws upon a robust sample that is proportionately representative of the England and Wales population at the national level. By using national scores, it is argued that the seriousness ratings account for weightings of the country as a whole.

Reliability: Judgement weightings can be applied to multiple units, people, places or time periods. Sentencing guidelines rarely change, but newly introduced crimes (e.g., coercive control) can take time to be added into the survey. Analyses exploring the effects of different labelling of crimes on seriousness judgements have not yet been conducted (Ignatans and Pease, 2015). It is probable that the weightings would need to be recalibrated with each annual publication of the CSEW, thus limiting the ability to conduct analyses over time.

Cost: Seriousness judgement data are available online and can be broadly calibrated with Home Office crime classifications at an aggregated level (e.g., violent crime). The practical application of these would not be problematic, but the exclusion of homicide/attempted homicide from the model would likely skew results in domestic abuse analysis. Implementing localised versions of the judgements would be costly.

Table 7. Suitability assessment: Victim seriousness judgements

Resolves conflict democratically	Demonstrates reliability	Cost-effective (including practicable)
WEAK	MODERATE	WEAK

11.4.3 The Home Office Economic and Social Costs of Crime tool

Methodology: Weights crimes using financial estimates (in pounds sterling) based on the anticipation of crime, consequences of crime, and response to crime. Evaluates both personal and commercial crime types at an aggregated level. The model disaggregates costs within categories, including separating out the costs of emotional and physical harm, which is the focus of this research.

Conflicting views: The available methodology (Heeks et al., 2018) does not explicitly state if or how conflicting views on levels of harm are managed. The methodology establishes the cost by multiplying the likelihood of the harm by the percentage reduction in quality of life, which is multiplied by the duration, which is multiplied by the value of a year of life in full health (the method is known as QALY). Victims are asked to estimate the amount of financial compensation they believe they would require to balance out the harm and inconvenience they have suffered. It is not clear from Heeks et al. (2018) how this method deals with outliers, or what the sample sizes are. Likelihood is calculated using the CSEW to determine prevalence. Quality of life impact is drawn from Salomon et al. (2015; reproduced in Heeks et al., 2018; appendix I) and duration of harm from Dolan et al. (2005). Cost is based on the Department of Health's value of life statistic from 2012. None of these elements explicitly deals with conflicting views.

Reliability: At an aggregate level, costs can be applied to multiple units. However, reliability is undermined by two major issues: firstly, the methodological complexity, particularly around individual circumstances, and secondly, the impact of inflation, which would necessitate at least annual adjustment to ensure ongoing reliability.

Cost: The data are freely available online, and simple to apply to aggregated Home Office crime classifications.

Table 8. Suitability assessment: Home Office Economic and Social Cost tool

Resolves conflict democratically	Demonstrates reliability	Cost-effective (including practicable)
WEAK	WEAK	STRONG

11.4.4 Office for National Statistics Crime Severity Score

Methodology: Weights crimes using the five-year average sentence length as actually issued by courts (taken from Ministry of Justice data). Translates non-custodial sentences such as fines to an equivalent value. Applied to all forms of crime documented by the Home Office classifications.

Conflicting views: Resolves conflicting views on the basis of the real-world determinations of magistrates and judges as well as statistical smoothing through averages. The ONS argues that this is a practical reflection of the will of parliament and the judgement of the judiciary.

Reliability: Can be used for any unit of analysis to which a crime is attributable (people, places or units of time). Actual sentences are subject to variation, but the use of five-year averages of sentences, a term which ONS may even increase, reduce this impact. Over longer-range timeframes, the index would need to be adjusted more frequently, but this does not present a potential problem here as our research analyses historical data. This may be problematic for more recent offence types which do not have five years of data from which to generate an average (coercive control is the prime example), but such instances are relatively few and always temporary. The main possible issue with reliability is the extent to which actual sentences, based on the comparatively small amount of recorded crime that makes it to sentencing at court, are an accurate reflection of harm. As Ashby (2018) identified, CSS weightings tend to be higher than the minimum guidelines, which is probably explained by judges giving consideration to aggravating factors and repeat offending, leading to the issuing of sentences that exceed the minimum. But, the tool also has some imbalances which are difficult to justify conceptually. For example, the rape of a female over the age of 16 is weighted at 2,890, but the rape of a male over the age of 16 at 2,930. Conversely, the rape of a female child under 13 is weighted at 3,229, compared to 2,535 for the rape of a male child under 13. All offences that differentiate between victims on the basis of gender are imbalanced to some extent, which could lead to gender bias in any subsequent analysis.

Cost: The data are freely available online, and at a high level of specificity, with around 300 Home Office crime classifications available.

Table 9. Suitability assessment: ONS Crime Severity Score

Resolves conflict democratically	Demonstrates reliability	Cost-effective (including practicable)
STRONG	MODERATE	STRONG

11.5 Summary

This chapter has laid out the history of the development of harm measurement tools, which are crucial to the analysis in several of the questions this research seeks to approach. These tools break down into four strands, three of which have tools that are viable for our purpose based on the criterion that they can be readily applied to police data in England and Wales. An assessment of these against the three criteria set out in Sherman, Neyroud and Neyroud (2016) identifies that the Cambridge Crime Harm Index and the ONS Crime Severity Score are the only instruments which pass each test. Table 10 summarises the final outcome of our assessments, in descending order of overall viability. Ashby (2018) concluded in his analysis of the two instruments, either could make a viable analytic tool but this analysis finds that, the Cambridge model scores higher on demonstration of reliability in two respects. Firstly, it evenly balances offences which are distinguishable by victim gender, as opposed to the ONS model, which has imbalance in each category that separates by gender. Secondly the CCHI offers more stability in its harm weightings over time. On the evidence assessed, this research proceeds with the use of the Cambridge Crime Harm Index as the most suitable instrument.

Table 10. Final viability assessment of harm measurement tools

Tool name	Author(s)	Open access?	Apply to England and Wales?	Apply to police datasets?	Conflict resolved democratically	Reliable measurements over time	Cost effective
Cambridge Crime Harm Index	Sherman, Neyroud and Neyroud (2016)	☑	☑	☑	STRONG	STRONG	STRONG
ONS Crime Severity Score	Office for National Statistics (2016b)	☑	☑	☑	STRONG	MODERATE	STRONG
Home Office Economic and Social Costs of Crime	Heeks et al. (2018)	☑	☑	☑	WEAK	WEAK	STRONG
CSEW Victim Seriousness Judgment	Ignatans and Pease (2016)	☑	☑	☑	WEAK	MODERATE	WEAK
Canadian Crime Severity Index	Wallace et al. (2009)	☒	☒	☒	Not assessed		
Sentencing Gravity Score	Ratcliffe (2015)	☑	☒	☑	Not assessed		
Severity Typology	Sellin and Wolfgang (1964)	☒	☒	☒	Not assessed		

12 Targeting Domestic Abuse: The Evidence

12.1 Chapter roadmap

This chapter outlines the established evidence in each of the main groups of research question set out in the Introduction chapter. The aim of this chapter is to provide readers with the background information they need to place the findings in context and understand the wider implications of how they contribute to the evidence base.

The chapter begins with a review of literature on repeat domestic abuse, which is a prerequisite factor for many of the factors this research seeks to explore. Without repeat abuse there can be no escalation, no concentration of harm, no serial offending and probably low potential for forecasting.

There is then an extensive summary of evidence on serial perpetrators, an area which has had limited coverage in domestic abuse research to date. This section of the chapter includes synopses of work on general typologies of domestic abuse offender to contextualise how serial offenders are situated within the bigger picture.

There follows a section on previous research into escalation, which has become a widely accepted phenomenon in domestic abuse practice with an apparently limited empirical basis. This is followed by a brief section on the concentration of harm, a subject for which has virtually no prior research.

The chapter concludes with an analysis of previous research into forecasting domestic abuse, and machine learning forecasting methods in criminal justice settings in general.

12.2 Repeat domestic abuse

While in many cases a domestic abuse event ends the relationship between partners, research suggests that domestic abuse can also become a repeated phenomenon and a reflection of a wider pattern of events (Walby, 2005; Stark, 2007, as cited in Robinson, 2016). Evidence tends to support this argument for both offenders and victims (Bland and Ariel, 2015; Chambers-McClellan, 2002; Feld and Straus, 1990; Sherman, 1992; Walby and Allen, 2004). This body of research suggests that some domestic batterers and abusers find it difficult to ‘break the cycle’, while certain victims of domestic abuse are similarly ‘trapped’ in abusive relationships. The causes of this persistency are unclear, and our ability to predict either

which relationships will immediately end, or which will persist in spite of abuse is not well refined. There are multiple psychological, environmental and economic factors that appear to play a part in these decisions (see for example Elisha, Idisis, Timor and Addad, 2010; Malach-Pines, 2002; Mintz, 1980), yet we do know, with hindsight, that repetitive domestic abuse does exist, even if we are unable to fully characterise the extent, scope, and nature of such cyclic violence.

Our understanding of the repeat abuse phenomenon is limited by the precision and accuracy of the data we are able to collect. For example, much of our knowledge is grounded in public records, such as official statistics collated by police forces. Traditionally, the external validity of these data has been challenged, and a plethora of evidence suggests that domestic abuse is underreported. When compared to victims' surveys, for example, the 'criminological gap' can be large (Gracia, 2004; Felson and Pare, 2005; Frieze and Brown, 1989; Pagelow, 1981), although as Chapter 9 highlighted, this gap may be shrinking. There are also cultural variables at play (Kasturirangan, Krishnan and Riger, 2004); some types of victims are more likely to report domestic abuse, while others are less likely to lodge a complaint against a family member (Felson and Pare, 2005). The way in which the police might handle such a report is also an issue, which some have referred to as 'secondary victimisation' given the lack of necessary sensitivity on the part of police officers (Barnish, 2004).

Several studies highlight both victims and offenders separately as being party to repeated abuse. Hester (2013) found that 83% of male domestic abuse offenders repeated their offences in a six-year follow-up period, while Smith, Flatley and Coleman (2010) found over three quarters of domestic abuse incidents to involve repeat victims (see also Barnham, Barnes and Sherman, 2017; Kerr, Whyte and Strang, 2017; Stark, 2007; Walby, 2005). Feld and Straus (1980) found similar levels of repeat cases in a family violence survey in the United States, and Walby and Allen (2004) identified high levels of repeat criminality within relationships in the 12 months preceding their survey, with females experiencing higher levels of abuse than males. Scholars tend to disagree on the number of crimes victims typically experience, however, with much conjecture surrounding precise numbers (Bland and Ariel, 2015; Giles-Sims, 1983; Okun, 1986; Straus, 1990).

However, the general prevalence and overall importance of recidivism in domestic abuse is demonstrated by a large body of research (see for example, Lloyd, Farrell and Pease,

1994). Broadly speaking, such studies consolidate the evidence that repeat offending in domestic abuse is widespread, and many early US domestic abuse studies focused on reducing recidivism. Sherman and Berk's (1984) trial on the impact of arrest in Minneapolis found repeat rates of between 13% and 26% within six months among its different cohorts. This seminal study spawned numerous replications (known as the Spousal Assault Replication Series) in other cities around the US. While the studies had mixed findings on the effect of arrest on repeat offending, they did all find widespread evidence of repeat offending (Maxwell, Garner and Fagan, 2002). While the Minneapolis trial and its replications focused on police reports as an outcome measure, others have used victim surveys, but all have reached a similar conclusion: repeat victimisation is common, lying somewhere between 17% and 59% (see Felson, Ackerman and Gallagher, 2005).

Other researchers have conducted longitudinal studies of criminal careers. Klein and Tobin (2008) reviewed 342 men who went before court for domestic violence crimes in Massachusetts in 1995 and 1996, studying their criminal histories until 2004. The results were again similar: 32% of these men committed further domestic abuse within a year of their index offence, and 60% did so within the full duration of the study. This pattern of higher recidivism with longer study periods is supported by Loinaz's (2014) study of 150 Spanish males imprisoned for domestic abuse, of whom 15% committed a further domestic offence initially, rising to 66% within a year.

Scandinavian studies have reported prevalence rates of between 16% and 48% for differing types of domestic offender (Svalin, Mellgren, Torstensson-Levander and Levander, 2017; Petterson and Strand, 2017), while British studies have provided an array of supporting evidence. Hester's aforementioned study (2013) tracked 96 domestic abuse offenders for six years and found 83% of male perpetrators to reoffended in that timeframe. This study was the latest in a series of papers by Hester focusing on one police force in northern England. Earlier research had found that half of offenders had a repeat domestic case within three years (Hester and Westmarland, 2006) and that those perpetrators described as 'all round offenders', or those who committed other non-domestic forms of criminality, were more likely to reoffend (Hester et al., 2006). More recently, Bland and Ariel (2015) analysed 36,000 police records of domestic abuse in Suffolk, in the east of England, and found that 35% of suspects were linked to more than one police-recorded report. Bland and Ariel's study differed from others in that it incorporated police records of non-crime incidents for the first time and used a non-judicial definition of the term 'offender'.

12.3 Serial Domestic Abuse

Research specifically on serial victims and offenders is less dense than research on general repeat abuse. The notion of a 'serial domestic abuser' is better known than that of a 'serial abuse victim', often driven by media attention and the wider expectations around serial criminals (Robinson, 2017). In this subsection we firstly consider how domestic abuse offenders have been typically classified in previous research. The subsection then expands specifically on serial offenders and what empirical research has so far concluded about their prevalence.

12.3.1 Typologies of domestic batterers

A wide body of research has attempted to develop a taxonomy of classifications for domestic abuse offenders. Among these studies, the evidence strongly indicates heterogeneity (Cantos and O'Leary, 2014; Cavanaugh and Gelles, 2005; Gondolf, 1998; Gottman et al., 1995; Hamburger, Lohr, Bonge and Tolin,1996; Holtzworth-Munro and Stuart, 1994; Johnson, 1995; Johnson and Ferraro, 2000), and from this basis a range of typologies have emerged, largely from the field of psychological research. Consequently, typologies tend to fall into two main groups: a small number of behavioural-based models and a large number of personality-based models. Behavioural-based typologies were explored by Brisson (1981) and Gondolf (1988). Gondolf identified three types of batterers: type I – 'sociopathic' abusers, who commit high levels of physical and social abuse; type II – 'antisocial' abusers, who are generally more violent but less likely to be arrested; and type III – 'typical abusers', who are generally violent but less disposed to serious violence. There are clear parallels between Gondolf's typology and Johnson's common couple violence/patriarchal terrorist taxonomy (Johnson, 1995), its subsequent expansion (Johnson and Ferraro, 2000), and the personality-based models developed by other researchers, as described by Cavanaugh and Gelles (2005) in their synthesis of the evidence.

Johnson and Ferraro's (2000) work identified typologies of relationships rather than offenders, but the taxonomy types still have relevance by way of implicitly describing the characteristics of perpetrators. 'Common couple violence' offenders are only violent within their relationships, and according to the authors, are approximately evenly split between males and females. In these couples, violence occurs just once or twice. This doesn't necessarily preclude 'common couple' offenders from being serial offenders, but this doesn't fit the narrative of 'predators stalking prey', and logically it may take a long period of time for offenders to accumulate multiple victims as they move through relationships. 'Intimate

terrorists', on the other hand, could potentially fit the serial stereotype well. This type of violence is explicitly relevant to the offender who uses violence as part of a wider pattern of control and coercion. Johnson and Ferraro suggested that these types of offenders are more dangerous to victims.

The third Johnson and Ferraro relationship type, 'violent resistance', explains offenders responding to a threat from a victim who is normally the aggressor. It is less logical that such offenders would be serial, although it is of course still possible.

The fourth type, 'mutual violent control', is an extension of intimate terrorism, but on the parts of both parties. The fifth and final kind of violent relationship defined by Johnson and Ferraro, 'generalist-borderline violence', is an extension of earlier attempts to define antisocial batterers (Holtzworth-Munro and Stuart, 1994; Jacobson and Gottman, 1998). In these relationships, the offender undertakes violent acts as a symptom of being emotionally overwhelmed.

Personality-based typologies first emerged with Elbow (1977), who described four personality syndromes in wife abusers, predicated on a combination of social learning and family perspectives. Elbow's types ('controller', 'defender', 'approval-seeker' and 'incorporator') have less validity today than at the time of their creation owing to social changes and the growing knowledge of domestic abuse outside the context of heterosexual marriage. However, the generalisability of types and integration of theoretical constructs certainly influenced later research.

Perhaps the most prominent typology to subsequently emerge is found in the work of Holtzworth-Munro and Stuart (1994), who reviewed 15 separate typologies to identify groups along theoretical lines of severity and frequency, generality of violence, and psychopathy and personality. This work led to the development of a tripartite typology: 'family-only abusers', who only commit crime in the domestic setting; 'generally violent abusers', who commit violent crime beyond their family and home; and 'generalist abusers', who straddle the other two groups. Holtzworth-Munro's model is somewhat simplistic to the point of over-generalisation, but it has a high level practical application that articulates a continuum of severity, and it has been validated to an extent by other research, including that of Johnson and Ferraro (2000), which offers multiple parallels. Hamberger and Hastings (1988), albeit with a very small sample (n = 204), found three clusters of spouse abuser type, which correlated with Holtzworth-Munro and Stuart. Saunders (1992), also with a small sample (n =

165) identified three types of abuser, too ('family only', 'generalised' and 'emotionally volatile').

Other models have been developed, too. Tweed and Dutton (1998) concluded that prior research pointed to two distinct subtypes of batterer: a group which supresses conflict in marriage and thus commits violence in non-intimate relationships, and a group that reports only intimate partner violence.

Among these numerous typologies there is one common theme: the classification of the extent to which the offender's violent behaviour pervades outside the domestic environment. This particular aspect is supported by other research (Klein, 1996; Buzawa, Hotaling, Klein and Byrnes, 1999), and there is a tangible, if abstract, agreement between the models that the frequency and severity of recidivism vary between typology, whatever the framework. Cavanaugh and Gelles (2004) summarise this effectively in their description of a different overarching trifold typology: low-, moderate- and high-risk offenders, for which they assert that little escalation occurs from low to high. This is perhaps the most relevant research for practitioners who are primarily focused on the management of (relatively) short-term risk of reoffending, and most likely to be recognised due to its similarity in descriptive structure to risk assessment cohorts.

Despite this array of typologies, there are large gaps in the evidence base for their application and theoretical design. Most research is based on married men, and little is known about the fit of models within different demographic groups, particularly minorities. Critical reviews of typologies have concluded support for the Holtzworth-Munro and Stuart framework (Dixon and Browne, 2003) and even suggested an expected proportion for each typology, but explicitly criticise research to date for a narrow focus on offenders only and the lack of a systematic approach to offender profiling. This problem is compounded by the relative ambiguity of definitions in this field. As Edelstein (2016) highlighted, researchers and scholars frequently misuse terms. Edelstein suggested that theory should form an agreed basis for constructs such as typologies. In the absence of this, this area of research, although rich in volume, is confusing, ambiguous and at times contradictory, limiting the potential for its meaningful practical application, at least in the law enforcement field.

On the basis of this limited practical usability it is unsurprising that typologies of domestic abuse offenders have gained little traction among offender programmes (Cantos, Goldstein, Brenner, O'Leary and Verborg, 2015). Given the current established practice in

England and Wales of using a 'high/medium/low' system prompted by the DASH process (see Chapter 9), any system typology designed for practical application must seek to augment or complement this design. Some researchers have suggested that categorising offenders on the basis of their violence profile would be a useful development (Petterson and Strand, 2017), while others have suggested this should be a prerequisite for assigning the correct intervention to an offender (Cavanaugh and Gelles, 2004). Whatever the option, there is general agreement about the need for such a tool (Petterson and Strand, 2017; Stoops, Bennett and Vincent, 2010; Edelstein, 2016; Cavanaugh and Gelles, 2004).

12.3.2 Serial perpetrators

Serial perpetrators are potentially one variation on such a typology framework, but evidence in this area in particular is thin. However, the term 'serial perpetrator' has gained some traction among practitioners in recent times, as demonstrated by the public statements of intent from Chief Officers (Robinson, 2017) and the police inspectorate (HMICFRS, 2014a, 2015; Robinson, 2017). The label first came to attention following Richards' (2004) review of almost 400 domestic homicide and sexual assault cases in the London area. Richards identified a number of serial offenders who went from relationship to relationship committing abuse. While important in promoting the concept of serial abuse, Richards confused the definition of serial domestic abuse with serial violence, arguably conflating two of the Holtzworth-Stuart and Munro subtypes into a single definition. The number of serial offenders was also unspecified, making it impossible to gauge prevalence. Nevertheless, the concept has taken off, with the author since leading a campaign in the media for a register of serial domestic abuse offenders similar to the sex offender register.

The question of definition is of particular prominence in this topic. There is no consensus on what makes an offender a 'serial offender', presenting a significant obstacle to both the development of knowledge and the application of practical solutions. Although police chiefs in England and Wales have developed their own definition (Robinson, 2017), this is unique to police in those countries. Overall, there have been a number of methodological variations in defining serial offenders (Kocsis, Cooksey and Irwin, 2002). The term 'serial' is determined based on differing considerations, primarily frequency of crime or motivation. For the Federal Bureau of Investigation, three homicides are required to render an offender a serial killer (Kocsis and Irwin, 1998). Elsewhere, serial rapists require just two victims (Hazelwood and Burgess, 1987). Other researchers have specified the inclusion of a minimum elapsed time period between offences or that offences must be of the

same category (Best and Luckenbill, 1996; Egger, 1985; Holmes and Deburger, 1998; Mitchell, 1997). Defining seriality this way is problematic for a number of reasons, not least the arbitrariness of setting a threshold (see the arguments put forward in Edelstein, 2016; Kocsis and Irwin, 1998), so a range of alternatives have been advanced based on psychological or motivational factors. Kocsis and Irwin (2009) proposed that a psychologically-based method could lead to the identification of a serial offender when they have committed only one offence, but their analysis is not specifically aimed at domestic abuse and discounts the limited operational practicability of such a model. Edelstein (2016) is an advocate of 'criminal careers' as the defining characteristic of serial offender status, proposing distinctions between the professional career criminal who is motivated by material profit, the serial pathological career criminal seeking to pathologically profit, and the serial non-professional, who lacks any professionalism and offends out of habit. This theoretical construct poses interesting questions for research into serial domestic abuse and has overt parallels to Johnson's (1995) and Holtzworth-Munro and Stuart's (1994) frameworks. The key challenge for practical application, however, is whether the desire for pathological profit can be determined.

12.3.3 Prevalence of serial perpetrators of domestic abuse

Issues with definitions notwithstanding, there is at least some evidence that serial perpetration of abuse occurs and is moderately prevalent. Hester and Westmarland (2006) conducted the first UK-based research in conjunction with the Home Office. They studied 692 domestic violence perpetrators from the north-east of England, 90% of whom were males, over the course of an 18-month period and found that 50% of offenders had at least one further domestic incident, with 18% of those involving a different victim. This gave an overall serial prevalence rate of 9%, just half the 18% later reported by Robinson (2017). Hester and Westmarland described serial perpetrators within the context of four groups of domestic abuse offender: a 'one-incident' group, for which their index offence was the only known record in the study period; a 'mainly non-domestic' group, who had one domestic offence and more than one other type of crime; a 'dedicated repeat domestic violence' group which committed multiple domestic crimes but no crimes of other kinds; and an 'all-round repeat offenders group' composed of those with numerous types of offences, both domestic and non-domestic. This latter group was marginally the most prevalent. Hester and Westmarland only made passing reference to serial perpetrators, however, and it is not known how they are distributed within the 'dedicated' and 'all-round' groups.

In the UK, there have been two other studies with notable findings on the prevalence of serial domestic abuse perpetrators. Firstly, Bland and Ariel (2015) analysed 18,675 offender cases from Suffolk between 2009 and 2014 using big data analytic methods. With data incorporating non-crime and crime incidents, they identified a repeat rate of 35% and, within this cohort, a 47.6% serial rate, giving an overall serial perpetrator prevalence of 16.7%. Secondly, Robinson (2017) analysed a range of police and partner data sources pertinent to 100 domestic abuse perpetrators in Wales. Obstructed by data quality, Robinson's estimate of prevalence among this cohort ranged from 4% to 20% owing to the disparity in definitions used by different agencies. The problems Robinson encountered offer stark insight into the difficulties of exploring the serial tendencies of offenders in a practical setting, but ultimately the generalisability of this research was limited by a small sample size and narrow geographic focus. Robinson concluded that serial considerations should form part of offender management decisions alongside risk assessments but did not develop this idea beyond a strategic overview.

Research into serial perpetrators of domestic abuse from outside the UK is even sparser, but generally finds higher rates. Klein et al. (2005) found that 28% of 552 male offenders on probation in Rhode Island offended against a different victim within a year. Bocko, Cicchetti, Lempicki and Powell (2004) found that 43% of 1,341 offenders charged with violating a restraining order had more than one victim. This work was limited, however, by the exclusion of a third of its original sample which did not have viable relationship information.

12.4 Escalation

The term 'Escalation' commonly refers to the phenomenon of increasing chronological severity, be it generalised or linear. It is a component of the current risk assessment model used by domestic abuse practitioners, and its roots can be found in theories developed more than three decades ago (Pagelow, 1981; Walker, 1979, 1984). It seems that, at least in the popular view, 'most calls to police or survivor advocacy agencies only occur after survivors have experienced lengthy escalation' (www.abuseandrelationships.org). Yet the evidence lays out a more complicated story, with fewer systematic observations of its existence than we might like to see to conclusively satisfy questions such as whether abuse tends to increase in severity over time, or is 'higher' harm either random or circumstantial? The answers to such questions pre-empt whether temporal patterns can be identified and, by implication,

predicted? Similarly, if an escalation in severity is predictable, what functional shape characterises its growth? The epistemological and phenomenological antecedents of domestic abuse have been studied over time, yet scholars do not agree fully on the evidence of these harm pathways. One clear example is domestic homicide, with the hypothesis that harm increases over time and culminates in a moment of killing, but the link between domestic abuse and domestic homicide is far from clear. Some research suggests that persistent domestic abuse and domestic homicide do not share similar characteristics, as the offenders' justifications for domestic homicide are often different than those offered by domestic abusers (Goussinsky and Yassour-Borochowitz, 2012). If this is the case, then it could be argued that escalation in harm – from verbal abuse and controlling partnerships to physical and ultimately homicidal victimisation – is rooted in discrete aetiologies (see for instance Moffitt, 1993, more broadly on differential crime growth taxonomies).

Given the different views on escalation of severity, it is not surprising that there is inconsistency in the evidence. Walker (1984) and Feld and Straus (1989) used surveys which relied on victim accounts and argued that escalation does take place. Campbell, Glass, Sharps, Laughon and Bloom (2007) concluded that violence by males against their partners is the most 'salient risk factor' for homicide, as domestic violence precedes up to 70% of cases. Similar conclusions were demonstrated by Crawford and Gartner (1992) as well as Stout (1993). However, Feld and Straus (1989) compared only two temporal data points, therefore lacking the necessary sensitivity and variation over time to demonstrate a developmental function.

On the other hand, Chambers-McLellan (2002) found perhaps the clearest evidence of escalation among 19,686 residential domestic abuse cases in Georgia, USA. The study concluded that crime severity increased by 0.07 on the Conflict Tactics Scale (a 0–18-point scale of severity) but had clear limitations, using a timeframe of only 12 months, with notable sample exclusions and an unequally weighted measurement instrument. Other researchers have failed to replicate the extent of findings (Piquero, Brame, Fagan and Moffitt, 2006 – though this studied used a truncated scale of severity measurement), leading Dutton and Kerry (2002) to conclude that domestic homicide does not necessarily follow escalation of violence. Given that homicide is relatively rare, this is not surprising, and the findings do not rule out the presence of escalation in more serious cases other than homicide. The problem then is classifying 'high harm' or serious cases other than homicide.

Other evidence to support at least partial escalation can be found. Johnson (2006) identified that just over three quarters of intimate terrorism cases indicated that violence became more severe over time. A similar proportion of participants in Andersen, Gillig, Sitaker, McCloskey, Malloy and Grigsby's 2003 study made a similar indication.

The lack of homogeneity across results is at least partially explained by the absence of a reliable measure of crime severity. Bland and Ariel (2015) attempted to address this by introducing the use of CCHI as a measurement instrument (see Sherman, Neyroud and Neyourd, 2016) which, as discussed in Chapter 10, provides a more robust way to assess harm over time. Yet this approach did not lead to observation of statistically significant patterns of escalation among domestic abuse dyads in Suffolk, England, with five or more reported incidents in a five-year period. The CCHI was also recently used to investigate escalation in domestic abuse cases by Barnham et al. (2017) in the analysis of 52,296 perpetrators of intimate partner violence in the Thames Valley police jurisdiction, and by Kerr et al. (2017), who analysed more than 60,000 records from the Australian Northern Territory. Neither study found evidence of escalation other than among Aboriginal offenders with three or more intimate partner incidents in a four-year period (Kerr et al., 2017).

From a policy perspective, escalation of violence has been assumed to be a risk factor for domestic homicide for some time (Campbell, 1995); however, a substantial number of abusive relationships are not known to the police or other social services (Aldarondo and Mederos, 2002). Therefore, it remains an open question whether the 'writing was indeed on the wall' of the police station and whether violence – and specifically domestic homicide – could have been prevented by predicting future harm based on past harm reported to the police.

12.5 Concentration of harm

Research concerning the extent to which harm in domestic abuse cases is concentrated, is limited to just two studies. Bland and Ariel (2015) used the Cambridge Crime Harm Index to conclude that less than 2% of all dyads in Suffolk, England accounted for 80% of cumulative harm, of which half of the dyads had no prior record of domestic abuse. In a partial replication of Bland and Ariel's study, Barnham et al. (2017) found that 3% of domestic abuse offenders in the Thames Valley police jurisdiction accounted for 90% of cumulative harm.

12.6 Forecasting

If we follow current practice and theory of escalation, high harm domestic abuse *can* be forecast before it occurs, based on prior behaviour. In Chapter 19 we will test this explicitly using a machine learning technique. In this subsection we critique the recent history of actuarial forecast tools in general and machine learning tools in particular, in criminal justice environments.

12.6.1 Actuarial instruments in criminal justice forecasts

Assessments of dangerousness in domestic abuse cases have existed for a long time, in a number of countries, resulting in a large number of varying instruments for the task. In the last two decades, the development of tools has accelerated as the demand on agencies charged with dealing with domestic abuse cases has increased. As one of the primary figures in domestic abuse danger assessments, Jacquelyn Campbell, explained in her Vollmer Award address (2005), these instruments offer a method of triage, essential to allocating limited resources efficiently. At the heart of this expansion, one key issue has remained consistent: should instruments be based on clinical methods (in which forecasts are arrived at by expert panels or individuals, based on professional judgment and experience) or actuarial methods (in which forecasts are derived from an empirical, often mathematical, basis)? In fact, there are three forms of model in practice using 'structured professional judgement', which comprises both clinical and actuarial elements. Each form has its exponents and critics, often based on interrelated aspects; a 'con' for actuarial is a 'pro' for clinical, and vice versa. Though well-worn, in the context of assessments of future dangerousness, these arguments are worthy of our attention.

The development of clinical instruments in danger assessments has been the product of practicality and cost more than of rigorous research. One argument holds that a victim's own perception of their future risk is as good a predictive tool as any, and indeed research offers limited support for this claim (Campbell, 2005). In the context of mental health and future violence, it has been argued that the 'science' is not available to design actuarial models of sufficient effectiveness to replace clinical judgements (Litwack and Schlesinger, 1999). Two decades later, this is not necessarily the case. Even before Litwack and Schlesinger's claim, scholars were arguing that actuarial models offered the only 'defensible' option (Quinsey et al., 1998), and indeed most empirical studies, even dating back to Paul Meehl's original focus on the actuarial versus clinical debate (1954), have in the main concluded in favour of the superior accuracy of the former. This is best summarised in two

meta-analyses of actuarial versus clinical model studies set out in the 2000s. Grove, Zald, Lebow, Snitz and Nelson (2000) found that actuarial techniques (described by the authors as 'mechanical') 'substantially outperformed clinical prediction in 33%-47% of studies examined' (p. 19). Conversely, just 6–16% of the 136 studies included in the analysis found the difference in predictive accuracy to be substantially in favour of clinical methods. Only six of these studies related to criminal recidivism or criminal behaviour, but the overall effect, and its order of magnitude – that actuarial instruments are more accurate than clinical tools with a Cohen's *d* of 0.12 – is of note in the debate overall, as is the authors' assertion that the superiority of actuarial tools is not universal.

Ægisdottir et al. (2006) built on Grove et al.'s work in the field of mental health practice in particular, also finding that actuarial instruments made more accurate predictions than their clinical counterparts. In the 48 most rigorous studies in the analysis, actuarial tools were 13% more accurate on average. However, Ægisdottir et al. highlighted a number of subtleties which are worthy of examination in light of our purpose. They identified that statistical rules should be established, particularly where errors were of differing costs, and that not all statistical tools were equally accurate, nor were they all more accurate than clinical tools. They also emphasised the need for practitioners to be familiar with any tools used, particularly in respect of their ethical implications.

While the consensus of researchers is that actuarial models are more accurate at predicting outcomes than clinical methods (see also Gottfredson and Moriarty, 2006; Hastie, Tibrishani and Friedman, 2009; Milner, Campbell and Messing, 2017), they are far from being the dominant form of tool used in practice. Researchers commonly agree that actuarial tools *can* be better at using data more reliably and consistently, paying regard to base rates, and allowing for more accurate profiling of weights. However, actuarial methods require statistical expertise to build and deploy and can be costly, so it is unsurprising that most police forces do not use actuarial assessments in the field of domestic abuse. Alive to the practical difficulties of deploying actuarial tools, researchers have recommended models based on combinations of clinical and actuarial methods, commonly labelled as 'structured professional judgement' tools (Kropp, 2004). The primary problem with this strand of tool is that it is even less specific than either of the others, and so potentially open to the problems of each. The devil, it seems, is in the detail that determines the precise role of factors which may influence the accuracy and fairness of forecasts (Urwin, 2016).

Foremost among researchers' concerns is the role that heuristics play in clinical or structured professional judgement–based forecasts. By definition, these assessments rely (at least in part) on procedures where clinicians gather and interpret information through the subjective lenses of experience, training, and their own world views (Meehl, 1954; Grove et al., 2000; Campbell, 2007; Robinson et al., 2016). Even if we accept that all the data gathered for interpretation in this way is consistent (which it is almost certainly not, in practice – see Robinson et al., 2016, for discussion of data gathering for domestic abuse by police in England and Wales), it remains inevitable that different people will arrive at different conclusions for identical cases. In this process, heuristics are integral. The potential for heuristic bias in forecasting was classified by Tversky and Kahneman (1975) and reviewed in respect of policing forecasts by Urwin (2016), who concluded that the array of possible heuristic factors influencing the decisions of custody officers was extensive. This framework is applicable to our research, in which individuals, including generalists and specialists, make assessments of future risk in domestic abuse cases. The individual's assessments are potentially affected by how readily they can recall relevant information, how many times they have encountered a similar scenario, and their confidence in their own knowledge (Tversky and Kahneman, 1975; Kahneman, 2011; Urwin, 2016). The latter is commonly overestimated, with even individuals who know that actuarial tools are generally more accurate preferring to 'go with their gut' on an individual case basis. High-profile, rare events can have undue influence on 'expert judgement' precisely because they are easier to recall (Kahneman and Klein, 2009). In theory, actuarial models can eliminate biases caused by heuristics, but this is not a given. If the predictor data on which an actuarial model is designed contains results that are the product of biases, these may trickle down to the resulting model (Harcourt, 2014).

In practice, the role of heuristics has not been definitively scrutinised in domestic abuse dangerousness assessments. In England and Wales, the setting of our research, the most prevalent form of this assessment for domestic abuse is the DASH, which is most commonly described as a structured professional judgement exercise. Though not applied consistently (Robinson et al., 2016), a common application of the DASH is as follows: a responding officer collates the answers to 25+ questions, asked of the victim. Each affirmative answer receives one 'point' and contributes to a total score, a method first established almost 100 years ago (Burgess, 1928). In all forces, this numerical score is combined with a professional's judgement of risk to determine the outcome. This process is hypothetically

repeated afresh each time a call-out is made to a domestic abuse incident. Robinson et al. (2016) conducted the most comprehensive review of DASH to date, conducting observations, interviews and surveys in three forces. Their review highlights several key points about the DASH risk assessment process as it was at the time, that are of relevance to any critique of structured professional judgement or the potential application of new methods. In concluding that the DASH was applied inconsistently by police officers, the authors explained that they found evidence of officers adjusting or omitting questions, or in some cases choosing not to submit a form at all. They also found that officers tended to weight criminal offences, in particular giving greater weight to those involving physical harm, and that attention to coercive control behaviours was missing.

The Robinson review recommended a more 'evidence-based' approach, and subsequently co-authors Julia Wire and Andrew Myhill evaluated the pilot of a new risk assessment (Wire and Myhill, 2018). The new risk assessment placed greater emphasis on coercive control and concluded that the tool led to higher rates of agreement between responding officers' and secondary risk assessors' judgements of risk. However, the methodology did not use equivalent comparison groups or test risk assessments for their predictive validity. The new tool increased the numbers of cases graded as 'medium' risk,[9] which has potential demand implications for police forces. At the time of writing, additional forces were piloting the new instrument with a view to nationwide roll-out, even though the question of how effective the new tool is at predicting high–harm domestic abuse had still not been addressed. Whether this is even an important question or not remains a matter of debate, but Campbell (2005) presented the most conclusive summary of important issues for domestic abuse risk assessment and tackled this issue in particular. Campbell argued that, before considering the issue of predictive validity, the agency using any domestic abuse risk instrument must first decide what it is for – to predict extreme violence such as homicide or simply the risk of reoffending. The latter is far more prevalent than the former in England and Wales, but is still relatively rare overall (Bland and Ariel, 2015; Barnham et al., 2017). It would seem that the DASH was conceived on the premise of the former (Richards et al., 2008; Robinson et al., 2016), but in either case the central issue remains one of prediction, so it is perhaps surprising that, in an area of such high demand and profile, neither the primary predictive instrument nor its proposed replacement has as yet undergone a countrywide or otherwise extensive assessment of predictive validity, especially when single-force research

[9] The typical domestic abuse grading structure is standard, medium or high risk.

(Thornton, 2017; Strang and Chalkley, 2017; Turner et al., 2019) has strongly indicated a tendency toward low predictive validity, particularly a high rate of false negatives. The latter of these three studies offers the most comprehensive view yet of the DASH's predictive validity. Turner et al., focussed on the 30 percent of a metropolitan police force's records (n = 350,000) in which the couple had more than one DASH record, and isolated those cases which the couple had no prior DASH record in the two preceding years resulting in a final sample of n = 61,080. Within this sample, they authors sought to establish the predictive validity of the 27 individual risk assessment questions and the overall risk grading (high, medium or standard). The outcome examined was the occurrence of future serious abuse (defined as assault with injury and above on the Crime Severity Score scale) within one year of the index crime. Among the cases that were re-victimised in this way, officers correctly risk assessed (i.e. gave a grading of 'high') in 5.7% of intimate partner cases (n = 41,570) and 2.7% of non-intimate partner cases (n = 19,510). By implication then, the false negative rates were over 90% for both categories. The false positive rates for officer predictions (i.e. cases where officers predicted some risk of future harm, but none occurred), was 94.4%. The authors concluded therefore that the DASH forecasts were not much better than random predictions, although they noted some caution because of the possibility that interventions in high risk cases may be responsible for some of the false positive results.

The implication of these findings concerning the predictive power of the DASH are especially contradictory to the current resourcing predicament in policing. If resources are tight, why are forces content to continue to potentially over-allocate resources to cases that will not result in either homicide or any form of reported reoffence? There is an ethical perspective to this issue. Not all abuse is reported, so it might be argued that the abuse which is reported merits investigation. But the practical aspects of this argument cannot be ignored; there are not enough police and partner resources to go around, and the risk of over-committing to cases unnecessarily is that those cases truly at risk of high harm do not receive the preventative treatment they require. One solution would be to allocate more resources to this area of business. Another option, and one that may be more efficient operationally and financially, would be to find a risk assessment instrument with high predictive validity.

12.6.2 Machine learning techniques

In this research, we consider the potential for actuarial instruments to fill the predictive void. In particular, we will examine machine learning techniques, a new branch of statistical method made possible by advances in computer processing capacity. Machine learning uses

computing to improve automatically (i.e., without human input at every stage) through 'experience'. Machine learning is used within artificial intelligence (AI) procedures and though they are often portrayed in the media as one and the same, machine learning is in fact distinct from AI. In general, machine learning comes in one of three varieties: supervised, unsupervised or reinforcement learning (Jordan and Mitchell, 2015). Supervised machine learning is the most commonly used form of the technique. It involves a human operator managing the algorithm(s) at every stage of the process by controlling inputs, reviewing outputs and adjusting accordingly. Unsupervised techniques have less human involvement in the control of inputs, often working with unlabelled or unstructured data. Reinforcement learning is a combination of the two techniques (see Jordan and Mitchell, 2015, for a full review).

Although the use of machine learning techniques for forecasting in criminal justice environments is relatively new, they have been enthusiastically adopted by some (Berk and Bleich, 2013). However, the emerging discipline has not been exempt from criticism, with some researchers contesting that the new instruments are no better than the old (Yang, Liu and Coid, 2010; Liu, Yang, Ramsay, Li and Coid, 2011; Tollenaar and van der Heijden, 2013) The primary premise of these criticisms is the condition of suitable transformation of data in order to allow the more traditional logistic regression methods to be effective. In contemporary times, this is a significant condition. While theoretically not unreasonable, in practice, it is extremely common for data to exist in unsuitable conditions. Berk and Bleich (2013) also contended that it is not logical that major international companies such as Google, Amazon and Microsoft would be employing new statistical techniques in their business models if they were no better than existing methods. At the root of this problem is the trend of considering new statistical developments as mere enhancements of the traditional linear models, whereas in practice they are not. Berk and Bleich highlighted a key distinction relevant to our consideration of domestic abuse forecasting:

> …a key distinction between forecasting and explanation has been badly conflated in some accounts (Andrews, Bonta and Wormith, 2006). Understanding a phenomena may lead to improved forecasting accuracy, or it may not, but forecasting and explanation are different enterprises that can work at cross purposes. (Berk and Bleich, 2013, p. 3)

The inference produced by this distinction has important implications. If accurate forecasts may be achieved based on more variables than only those with apparent correlative or explanatory relationships to the outcomes we seek to predict, then a multitude of additional data sources become available to us. The subsequent questions are (1) what methods may we use to seize such an opportunity? and (2) what degree of accuracy could such methods achieve? Maximising the accuracy of the forecasts should be a primary goal of forecasting instruments in criminal justice settings, Berk and Bleich asserted, because the resulting decisions have real consequences for people's lives. They concluded that adaptive machine learning techniques offer a superior alternative to logistic models owing to their ability to detect complex, non-linear patterns in datasets and set out a thorough framework for measuring forecasting tools against each other, comprising of (1) thorough establishment of what features are being compared, (2) comparisons based on data not used in the construction of the model, (3) appropriate comparison methods, (4) accurate characterisation, (5) comparable tuning parameter use and (6) close attention to practical interpretation. Using this framework, they compared the traditional logistic regression technique to two machine learning techniques – random forests and stochastic gradient boosting – and concluded that random forests offer the strongest and most flexible option. Random forests, an ensemble of classification trees (explained further in Chapter 14), offer all the primary benefits of machine learning methods, as described by numerous authors in recent times (Barnes and Hyatt, 2012; Berk, 2012; Breiman, 2001). Random forests, unlike other forms of machine learning, are not limited to the forecasting of binary outcomes such as 'yes or no'. They offer the ability to account for asymmetric costs such as in the case of criminology where serious crimes potentially have greater costs than less serious crimes. They build regularisation into their core calculations and can cope with a vast number of predictor variables, potentially making good use of the vast amount of data held by police forces. Importantly, they can cope with imbalanced distributions, for example where events (such as homicides) are rare, whereas traditional tools such as linear regression work most effectively when the distribution is simpler.

12.6.3 Previous use of random forests for criminal justice forecasting

The random forests technique has been used to construct criminal justice forecasts on several occasions in recent times. A summary of the examples of its use is worthy of consideration in the preparation of the forecasting methodology set out in Chapter 6, so the following paragraphs summarise the main studies to have used the technique, highlighting the context, methodological application and predictive validity in each case.

The first study to examine the random forests technique was that of Berk, He and Sorenson (2005), which attempted to develop a practical forecasting tool for the screening of domestic abuse incidents for the Los Angeles County Sheriff's Department. The authors of the study, particularly Richard Berk, would go on to contribute significantly to the body of random forest forecasting research in subsequent years, and this study, which tested the use of a single CART (Classification and Regression Tree) method and random forests (a multiple CART method), was a primer for the studies to come. The authors collected data on potential predictors from 500 Los Angeles households to which officers were called out and used a small subset of these to build a screening tool which they retrospectively tested against known outcomes. The objective of the forecasting tool was to predict future instances of any kind of domestic abuse at households within three months of the forecast, but the study was beset by practical problems. The intention was to sample a large range of houses with both prior and no prior domestic abuse records, but implementation failed in this respect, and the final sample was heavily skewed toward houses with prior domestic records. Officers also failed to ask all of the predictor questions required, resulting in listwise deletion being employed to deal with missing values in the data. Still, the CART model used was initially successful at identifying 66% of households with any new call for police service. The authors were concerned about overfitting (Breiman, 2001). Overfitting is the term given when statistical models are too closely aligned to a limited set of data points. When exposed to the 'real world' an overfitted model will not replicate its testing performance. To address this concern the authors tested the random forests technique as an alternative. This technique achieved 59% accuracy, but by using 'out-of-bag' testing, a process whereby a portion of the dataset is held back from model training to be instead used for validating the model, the authors concluded this to be a far more robust and reliable instrument. In relation to domestic abuse cases, the two techniques were also approximately equal in forecasting accuracy. Berk, He and Sorenson's paper contained many of the analytical points which became the hallmarks of later forecasting papers, including the use of confusion matrices to display the

models results, the overt consideration of cost ratios between false negative (predictions of no domestic abuse that were wrong) and false positive (predictions of domestic abuse that were wrong) errors, and consideration of the impact of individual predictors.

In Berk, Kriegler and Baek (2006), the same technique comparisons were made as in Berk, Sorenson and He (2005), but this time to forecast which prisoners were likely to commit serious misconduct while in prison. This outcome was found to be generally rare in the studied population, inmates in California, with fewer than 3% committing serious misconduct in a two-year period. Following the framework for forecasting analyses set out in the same year by Gottfredson and Moriarty (2006), the authors retained 1,000 of their overall sample of 9,662 for the purposes of testing their models. They then built forecasting models using logistic regression and CART techniques but found no notable improvement on the original marginal probability of 0.03. As with domestic abuse in Berk, He and Sorenson, 2005, the cost ratio was set to (1) one false negative (an offender being incorrectly forecast as committing no misconduct) having the same cost as ten false positives (an offender being incorrectly forecast as committing misconduct), and (2) one false negative to five false positives. The results indicated that random forests produced more accurate forecasts, correctly forecasting 49% and 62%, respectively, of misconduct for the two cost ratios. Their analysis also highlighted several key predictor variables which enhanced the accuracy of the forecasts.

Berk, Sherman, Barnes, Kurtz and Ahlman (2009) broadly replicated the methodology of the two earlier papers on which Berk had led. Their forecasting objective was the prediction of murder among a population of probationers and parolees. This diversion to the most serious form of crime marked a step towards attempting to forecast extremely rare outcomes, and the authors emphasised the 'high stakes' element of this matter in their title. A common theme in Berk's work is the attention paid to the practical and personal implications of the forecasts in question, which in this paper takes the form of a stark contrast drawn between the actual cost of a false negative (a homicide) and a false positive (wrongfully extended incarceration and overcrowding in prisons). The authors observed that forecasts will never be perfect, and so it is essential to pay attention to the balance of errors. In this respect, the paper counters arguments that the world would be a better place without statistical forecasting on the basis that their tool of choice (random forests) allows a structured process for the balance of errors to be accounted for whereas, in processes relying only on the subjectivity of individuals, no such overarching consideration can take place.

The study's target population was 60,000 cases from Philadelphia's Adult Probation and Parole Department. The objective of the forecasting model was to predict the occurrence of a homicide or attempted homicide within two years of the beginning of community supervision. The authors again took care to explain (as in the two papers covered previously in this section) that predictor values require no causal link to the outcome object of the forecast, but they did highlight the practical importance of establishing a form of 'common sense' link to add to a sense of legitimacy among staff using the tool in practice. They also carefully considered the use of only information that officers would have readily available at the time they needed to run the forecast.

The results of the random forest modelling were again set against the context of logistic regression performance. The latter resulted in a 99.7% error rate for the prediction of homicide. Using the same predictor information, random forests achieved a 57% error rate with little variation in performance when positive to negative cost ratios were adjusted between 7:1 and 12:1. When applied to test data, the model showed no indication of overfitting. Through the use of 'importance plots' (which demonstrate the contribution to overall predictive validity of each predictor variable) and 'partial response functions' (the pattern of each predictor variable's predictive validity), the authors also identified a small number of individual predictor variables as contributing substantially to the overall performance of the model, *inter alia*, age, age at first contact and the number of prior gun crimes. A version of the forecasting model was later used in a field experiment relating to supervision levels, in which it was employed to determine 'low risk' offenders as candidates for participation (Berk, Barnes, Kurtz and Ahlman, 2010).

In 2012, Richard Berk published a Springer Brief in Computer Science entitled *Criminal Justice Forecasts of Risk: A Machine Learning Approach*, which presented a detailed treatise on the justification and methodology for applying random forest modelling to criminal justice forecasts. In the book, Berk drew on the example of Barnes and Hyatt's (2012) work with the Philadelphia Adult Probation and Parole Department. This work expanded on the previous works published with the Philadelphia Department in great detail, developing a thorough 'dos and don'ts' approach to the design and implementation of a random forest model. It also tracked model performance through various iterations, each updated as new data became available. Like Berk, Barnes, Kurtz and Ahlman (2009), Barnes and Hyatt tested models not only on out-of-bag data but also against a totally independent 'test dataset'. Their recommendations for implementation included notes about data access,

outcome definition, predictor selection, cost ratios, tuning (which means the adjustment of sample sizes and other parameters which may change the performance of the model), validation and practical use in the field. These later informed the implementation in Urwin (2016) and the principles for responsible algorithm use set out in Oswald, Grace, Urwin and Barnes (2018). Barnes and Hyatt paid particular attention to the potentially controversial selection of some predictor variables, such as offender ethnicity, and their implications for the legitimacy of such forecasts. Furthering previous discussions on this topic, their work highlights an important issue in forecasting, which Berk (2012) also emphasises: forecasting accuracy is not the 'be all and end all' of model performance. A model has to be politically acceptable and operationally viable to stand a chance at successful implementation.

In the same year, Berk, Sorenson and Barnes (2012) published their development of a random forest forecasting instrument for domestic violence arraignment cases. Their objective was to determine whether a tool could be developed to usefully forecast the future dangerousness of domestic abuse offenders which may enable decision-makers to be better informed when deciding to release offenders or otherwise. They determined three outcomes for their model to forecast: (1) no arrests for domestic violence within two years, (2) a domestic violence arrest with no physical injury, within two years, and (3) a domestic violence arrest with physical injury, also within two years. Against a baseline situation of around 80% of those actually released at arraignment not being arrested for domestic violence within two years, the authors' model correctly predicted no arrest 90% of the time, leading to the general conclusion that, if magistrates used the model, they could improve the failure rate of decisions by around half. By virtue of cost ratios, the model predicted the other two outcomes less efficiently, over-compensating its forecasts to avoid a high false negative rate. Accordingly, while 74% of all domestic violence with injury was correctly forecast, only 21% of the total forecasts made for that outcome turned out to be correct. This is a critical point which we will return to later when assessing the performance of our own model.

The authors emphasised the use of readily available data, in effect recycling known information in a more efficient way by use of machine learning. Their analysis of 28,646 cases considered around 30 predictor variables, predominantly relating to an offender's criminal history. Age and gender were the only personal characteristics included; ethnicity was excluded. The analysis also included segments on the relative importance of individual predictors to the overall performance of forecasts, but the authors made no attempt to refine the model, arguing that even a small boost to predictive power was relevant. This study

represents the first published attempt to establish an algorithmic approach to domestic abuse forecasting, which we attempt to replicate in this Chapter 19. In this respect, many aspects of Berk et al.'s findings are discussed further in the Chapter 20 in relation our own findings.

Until 2016, every published paper on the use of random forest modelling for criminal justice forecasts involved law enforcement agencies from the United States. In 2016, Sheena Urwin, a police staff professional studying in the Cambridge Police Executive Programme, working with Geoffrey Barnes, wrote a thesis on the development and application of such an instrument in Durham, England (Urwin, 2016). The Durham Harm Assessment Risk Tool (HART) aimed to forecast the future dangerousness of arrested offenders presenting at custody suites. The composition of the forecasts was similar to those of Berk, Sorenson and Barnes (2012) and Barnes and Hyatt (2012) in that it presented three possible outcomes: no offence, any less-serious offence, or a serious offence; all within a two-year follow-up timeframe from when the forecast was made. The composition of the forecasting model followed the by-now-standard methodology for such tools. The model was trained and validated on separate datasets, with an 8% drop in overall accuracy and a 20% drop in dangerous forecast (false negative) accuracy, the latter of which prompted Urwin to emphasise the importance of regularly refreshing the model construction to account for changes in the operating context.

Urwin's HART model was conceived as the gateway triage tool to a deferred charge intervention. Offenders forecast as moderate risk (any less-serious offence within two years) would be eligible for a scheme known as Checkpoint in which they would be diverted from the normal criminal justice system on the proviso that they comply with specified conditions. Consequently, the balance of dangerous (incorrect forecasts of low risk) to cautious (incorrect forecasts of high risk) errors differed substantially from predecessor models because of the need to balance accuracy with the capacity of the Checkpoint programme. Previous random forest models (see Berk, Sorenson and He, 2005; Berk, Kriegler and Baek, 2006) employed a ratio of ten cautious errors (false positive) for every dangerous error (false negative). Urwin's HART model used a ratio closer to 3:1. Conversely, the probability of a 'very dangerous' error (a serious re-offender forecast as having no risk) was just 2%, and this, Urwin argued, was sufficient to enable the Durham Chief Constable to have sufficient confidence in the model to put it into operation. At the same time, by effectively deliberately over-forecasting the risk of individuals, Urwin did not eschew the ethical dilemmas her algorithm produced. The Durham HART model was widely scrutinised in the British media (Baraniuk, 2017;

Burgess, 2018), both in this respect and in terms of its fairness. Later revisions of the model removed the use of a sociodemographic classifier provided by the private company Experian due to complaints that it biased the model against less-affluent communities.

Urwin's paper was also the first random forest analysis to test the model performance against clinical judgements. Urwin developed a test wherein custody sergeants were given cases that were also tested by the model. The subsequent level of agreement was then analysed. In general, the predictions of the two methods were notably different. In moderate-risk cases, police officers and the algorithm agreed around two thirds of the time, but this dropped to around half the time in low-risk cases and less than a quarter of the time in high-risk cases. Urwin concluded that, typically, police officers were more risk-averse than the algorithm when it came to risk forecasting.

12.6.4 Criticisms and problems

Promising though the use of random forests for criminal justice and even domestic abuse forecasts may appear to be, as a branch of actuarial instruments, and in particular as a machine learning technique, the method is subject to the same criticisms and problems as other instruments of the same type. At the superficial level, these can manifest in popular media as 'big brother' issues, in which a person's hidden data play a disproportionate role in determining the consequences of their actions or the services available to them. In particular, critics have focused on the ethical and discriminatory aspects of particular variables, as was especially the case in response to the Durham HART model (Urwin, 2016; Baraniuk, 2017, Liberty, 2019). As this research proposes a replication of random forest forecasting, it is worthwhile summarising the main criticisms against and problems of actuarial criminal justice forecasting instruments identified by scholars and commentators thus far.

Perhaps the most comprehensive recent discussion of issues facing actuarial instruments was presented by Gottfredson and Moriarty (2006), who argued that the promise of such tools was as yet unrealised because key assumptions were being ignored or contradicted. Reviewing the original work of Gottfredson and Gottfredson (1986), the authors summarised the main issues surrounding implementation of actuarial tools as (1) the use of unreliable data in tools, (2) failure to consider the base rate (base rate meaning the rate at which the outcomes to be forecast occur within the population), and (3) the incorrect application of weighting factors. They also highlighted a number of potential methodological concerns which we return to in Chapter 20. These include the establishment of a cross-

validation sample, separate from the training dataset, the selection of appropriate measures of predictive accuracy, the consideration of static and dynamic variables and the inclusion of 'administrative overrides' for practical and ethical considerations.

The issues raised by Gottfredson and Moriarty have been frequently touched upon in publications concerned with criminal justice forecasts and random forests in particular (see Berk, 2008; Berk et al., 2009; Berk et al., 2010; Berk, 2011; Berk and Bleich, 2013). Berk and Hyatt (2015) brought together many of the ideas from those papers in their response to 'misinformed views' regarding actuarial models. They focused on five main criticisms: the legitimacy of actuarial instruments, insufficient levels of predictive accuracy, the double counting of predictor variables, the inability of models to be dynamic and respond to changing circumstances, and the potential for the introduction or consolidation of racial biases in decision-making. The authors concluded that clinical models, although appealingly simple to implement, also suffer from these issues, and drew attention to strategies to overcome each problem in actuarial models.

A primary issue of concern, specifically with random forests, is the 'black box' nature of the procedure, which refers to the 'unknowable' aspect of its calculations. In practice, a random forests algorithm calculates so many decision points (sometimes millions) that it is practically impossible to audit each forecast. This can leave practitioners uncomfortable, and critics claim that police do not really understand the decisions they are making (Berk, Sorenson and Barnes, 2012; Berk and Hyatt, 2015). The size of the calculations can also create practical implementation issues with securing the appropriate computer processing power or specialist software required to perform a forecast in a timeframe that does not inhibit police personnel from carrying out their duties (Barnes and Hyatt, 2012). This latter issue has been highlighted as one of the key considerations for law enforcement agencies when attempting to integrate machine learning techniques into their practices (Ridgeway, 2013).

12.7 Summary of the evidence

The literature summarised in this chapter presents a range of evidence for the existence of repeat domestic abuse as an important aspect of domestic abuse. Yet there is relatively little evidence for the trend of chronological escalation in severity. There are also a number of gaps in our understanding of serial perpetrators which presents a major obstruction to professional aims to target serial perpetrators. Foremost among these obstructions is the lack of an agreed

definition, but this is relatively simple to propose. More pressing is the need to obtain a more robust estimate of the prevalence of serial offenders, and then to describe and understand this subset of domestic abusers, particularly in comparison to other groups. Practitioners would likely find any additions to the evidence base in these respects helpful to their efforts to develop programmes targeting domestic abuse perpetrators. The literature indicates that serial perpetrators are a distinct group, albeit not in the majority. There is little to no understanding, though, whatever the size of the group, of their relative harmfulness. Calls for a register of these individuals – by means of which they would be tracked for life – would reflect a highly dangerous group of 'predators', but the potential impact of such a register has not yet been properly established.

Assessments of future dangerousness in criminal justice settings have been discussed for decades, with many iterations of the task evolving. Currently, forecasting instruments can be categorised into three classifications: clinical, actuarial, or structured professional judgement, with the last of these being the most commonly used in contemporary domestic abuse practice. In both clinical and structured professional judgement tools, heuristics are particularly important influencers of forecasting outcomes and one of the reasons that many scholars argue that actuarial instruments could improve on them. In England and Wales in particular, evidence shows that there is inconsistent application of the present tools, and there is no evidence at all regarding their predictive validity with reference to future dangerousness. Modern statistical techniques, combined with the large datasets now available to police agencies, open the possibility of using machine learning techniques such as random forests. Several studies, predominantly involving Professor Richard Berk and law enforcement agencies in the USA, have successfully established forecasting tools based on this model, but so far only one study has examined domestic abuse forecasts (Berk, Sorenson and Barnes, 2012), and only one study has examined the use of random forests in England and Wales (Urwin, 2016). A number of issues exist which the implementation of machine-learning-based actuarial instruments must address to stand a chance of being successful. These include establishing a clear framework for legitimacy and legality, including an assessment of ethics, thorough validation, consideration of base rates and appropriate IT design to enable practical use by frontline practitioners.

13 Research Questions

13.1 Chapter roadmap

In this chapter we revisit the research questions first set out in the Introduction, to place them in the context of the targeting methodologies reviewed in the Chapter 11 and set the scene for the methodologies we discuss in the next chapter. Each of the five primary research areas is covered in this chapter, which concludes with an overall summary that restates the overall direction of this research.

13.2 Repeat abuse

Table 11: Research questions: repeat abuse

Number	Question
1	What is the prevalence and extent of repeat victimisation of domestic abuse?
2	What is the conditional probability of further domestic abuse associated with each consecutive victimisation?
3	What is the prevalence and extent of repeat offending of domestic abuse?
4	What is the conditional probability of further domestic abuse associated with each consecutive offence?

Chapter 12 outlined that there is a large body of evidence that supports the existence of repeat domestic abuse. However. only a handful of these studies have examined police records as a primary source. Those that have, found repeat abuse to be prevalent in the minority of cases but in sufficient proportions to be influential on overall harm (Bland and Ariel, 2015; Barnham et al., 2017; Kerr et al., 2017). In practice, much emphasis is placed on targeting and preventing repeat abuse, so this warrants further examination. To this end, we focus on two primary research questions, subdivided by both victims (questions 1 and 2) and offenders (questions 3 and 4). The first and third of these questions deal with the issue of prevalence – what proportion of overall victims and offenders are linked to multiple cases, regardless of the identity of the other party involved? Based on previous studies of English and Welsh police records, we might expect this to be in the order of 25% of cases (Bland and Ariel, 2015; Barnham et al., 2017).

The second and fourth questions are predicated on the existence of repeat abuse and concern the probability of further repeats at each consecutive crime reported. The answers to the prevalence questions will reveal what the probability is of any domestic abuse offender or victim being linked to an additional domestic abuse crime. Conditional probability calculations will assess the probability of further crimes at each total level of prior abuse. Prior research of this nature is scant, but what exists has shown a generally increasing trend in probability with each additional crime report (Bland and Ariel, 2015).

Understanding the overall prevalence of repeat abuse is an important fundamental aspect to the research questions that follow, but they are also an important factor in our overall understanding of domestic abuse problems. If we find that repeat abuse over an extended period of time is in the minority we may place additional emphasis on the characteristics of this group or conversely, identify pertinent questions about those which do not experience repeat offending. Addressing these questions in a number of jurisdictions of varying geographic and demographic profiles, may enable more precise responses based purely on the prior number of domestic abuse crimes.

13.3 Serial abuse

Table 12: Research questions: serial abuse

Number	Question
5	What is the prevalence and extent of serial abuse among victims of domestic abuse?
6	What is the prevalence and extent of serial abuse among offenders of domestic abuse?
7	Are serial perpetrators demographically different from repeat offenders or single-time offenders?
8	What types of domestic abuse crime do serial perpetrators commit and how harmful are they?
9	Do serial offenders cause more domestic abuse harm than repeat or single-time domestic offenders?
10	To what extent do domestic abuse serial perpetrators commit other forms of crime, and how does this compare with repeat or single-time domestic offenders?

Evidence about the prevalence of serial offending is far rarer than repeat domestic abuse offending. As far as we can tell, it is non-existent for serial victimisation. In light of this, our first research questions are simple – what is the prevalence of serial cases among domestic abuse victims and separately, offenders?

Our subsequent questions then focus on perpetrators, which are, at present, the group which police and partners have an interest in, and for which more data are traditionally collected. Given the interest in targeting serial offenders, we will seek to establish a profile of the group relative to other 'types' of domestic abuse offender. It is of interest to determine if serial perpetrators have different demographic and offending history characteristics, and in particular, whether they truly are more harmful. Establishing the answers to these questions may reveal new avenues for agencies targeting serial offenders or it may call the whole endeavour into question. Either way, the answers are likely to be useful to practitioners and scholars, the latter of whom have not previously studied this group in any substantial detail.

13.4 Escalation

Table 13: Research questions: Escalation

Number	Question
11	Is there evidence of escalating harm in each consecutive domestic victimisation?
12	Is there evidence of escalating harm in each consecutive domestic offence committed by offenders?

With empirical evidence for the existence of escalating severity in domestic abuse crimes so sparse, there is real value to be gained from further measurement of the issue. While dyads have been the principal unit of interest to the few researchers who have explored escalation (see Bland and Ariel, 2015 as an example), the two research questions we will analyse relate to victims and offenders separately for two reasons. Firstly, this subcategorisation has not previously been explored, and secondly because some prevalence for serial victims and offenders is anticipated, we require a question that will account for such cases.

Understanding more about escalation, including whether it even exists in police records, has potential implications for both practical and theoretical agendas. In practice, risk assessments are partially informed by views on escalation. In theory, it is an established

premise (Pagelow, 1981, Richards et al., 2008; Walker, 1979). If escalation cannot be found, then this finding may challenge these positions. Alternatively, it may endorse them. In either scenario, these questions are potentially highly useful to the field.

13.5 Concentration of harm

Table 14: Research questions: Concentration of harm

Number	Question
13	What is the extent of concentration of harm among the most harmed victims of domestic abuse?
14	What is the extent of concentration of harm among the most harmful offenders of domestic abuse?
15	To what extent do the police have prior knowledge of the group of victims suffering the most harm?
16	To what extent do the police have prior knowledge of the group of offenders committing the most harm?

Previous research into concentrations of harm in domestic abuse cases is almost non-existent, so the four questions shown in Table 14 break new ground. Kerr et al., (2017) and Barnham et al., found evidence of concentrated harm using variations of the Cambridge Crime Harm Index, but these are isolated examples which require replication. Similarly, Bland and Ariel's (2015) conclusion about the half of 'high harm' cases having no prior records for domestic abuse merits further exploration. If this feature of domestic abuse harm were replicated more generally it would have considerable implications for domestic abuse strategies such as potentially limiting the preventative scope of risk assessments which, as discussed in Chapter 9, are typically only instigated after an initial report of domestic abuse. If half of high harm cases have no initial report, risk assessments will not take place until the high harm has occurred, meaning that at best, police could only hope to prevent half of their serious domestic abuse crimes before they happen.

13.6 Forecasting

Table 15: Research questions: Forecasting

Number	Question
17	What proportion of all arrestees go on to commit domestic abuse within two years?
18	What proportion of serious domestic abuse arrestees have prior arrest records in the two years preceding years?
19	Can antecedent inputs predict future domestic abuse cases to a high degree of accuracy?
20	Which inputs have the greatest impact on predictive validity?

If the few previous findings on concentration of harm are replicated by the analysis of questions 12-16, and police forces only have a chance of identifying and preventing around half of serious domestic abuse crimes, then it follows that there is a need to explore wider methods of predicting those serious crimes which occur 'out of the blue'. But, even if it were deemed morally and legally acceptable, population-wide screening is likely to be impracticable. Police forces would be better served examining other records they keep already, such as non-domestic abuse crime records. To this end, our research questions on forecasting focus on arrests for *any* form of crime, with a view to assessing the possibility of predicting domestic abuse (serious and less-serious) at the point a suspect is detained and processed through a custody suite. Questions 17 and 18 establish a baseline – what the prevalence of future domestic abuse within this population is, and of those arrested for domestic crimes, what proportion had prior arrest records – and so a chance of being identified by this method at all. The major question then follows: can a statistical model, which is based on prior criminal activity of an individual at the time they are arrested, successfully forecast future domestic abuse. If we can find such a model then it may have ramifications for crime prevention, police resources and even public health outcomes. But such a model would rightly be closely scrutinised to ensure it is not biased or compromised. These are major contemporary concerns with the use of algorithms in criminal justice (see Liberty, 2019; Oswald, Grace, Urwin and Barnes, 2018). The key challenge for algorithms is not just to be accurate, but to achieve that accuracy in a way that can be plainly and transparently understood as fair and just by an audience wider than just a few technical specialists. These factors necessitate examination not just of the predictive performance but

the inner mechanics of modelling too. So the final question posed is conditional on a 'successful' model being found – what are the predictor variables contribute the most to the predictive accuracy of the model? Understanding this detail may help address concerns about the legitimacy of the model or help future researchers refine it.

13.7 Summary

All of the research questions detailed in this chapter are directed at targeting domestic abuse. They build on previous empirical research which has shown that repeat offending and victimisation is common in domestic crimes, that to some extent serial perpetrators exist, that a small proportion of people are linked to most of the serious harm and that escalation of severity across case history is not necessarily known to the police. In setting these questions, the aim is to contribute to this body of research and identify strong evidence that can support the development of responses which single out the most harmful cases before they occur. The next chapter turns to how these questions can be answered.

14 Research Methods

14.1 Chapter roadmap

This chapter expands on the research questions detailed in Chapter 13 with the details of how the findings for each question have been produced. The chapter begins with a description of three datasets obtained for this research, then addresses each of the five research themes in turn. The first subsection explains the differences in the three datasets and which research questions they were obtained to address. Each was obtained at a different point in the research process, and though they all pertain to police domestic abuse records, each dataset features different variables chosen for the specific questions under examination. The first subsection outlines the nature of the police forces from which the data was obtained and specifically what variables the data include.

The following subsections then consider the analytical procedures we have used to address each research question. These include descriptive statistics, probability calculations and difference testing. The primary aim of this chapter is to support researchers looking to replicate this work, so each category is dealt with in some detail, with specific analytic methods outlined where appropriate. In the case of the statistical forecasting model we have developed, this chapter goes into a higher level of detail, including descriptions of how the chosen algorithm works, the key terms that describe its various parameters and functions, and how the model had been calibrated.

14.2 Three datasets

Three sets of police records were collated in total in order to address all 20 research questions outlined in Chapter 13. There were three primary reasons for compiling three datasets rather than one. Firstly, each category of question required different variables or parameters for its analysis. For example, to ascertain the non-domestic abuse crime harm of serial perpetrators (as required by Question 10), we needed to obtain details of non-domestic abuse offending history – a much larger set of data than just domestic abuse crimes. We did not need the same information to answer questions on concentration of harm or escalation, so were able to refine request to police forces to assist with streamlining our requests and reducing the burden on the agencies supplying us with the information.

Secondly, to increase the generalisability of our findings, data were requested from five separate police forces, from all over England and Wales, including rural and

metropolitan jurisdictions. How do five departments' data translate into three datasets? Dataset 1 comprised of four separate police forces' data. One of these forces also supplied records for Dataset 2 and a fifth, entirely separate force supplied Dataset 3. All details of where the datasets originated from have been removed from this work to comply with the conditions of data handling specified by the police forces supplying the data. The datasets were essentially a convenience sample. Enquiries were made with multiple forces and those responding positively were used to form the datasets.

Thirdly, each dataset was obtained at a different stage as this study progressed. Accordingly, the datasets correspond to different periods of time, which are specified throughout. The following subsections introduce each of the datasets for context and to assist those looking to replicate these analyses.

14.2.1 Dataset 1: Repeat abuse, escalation and concentration of Harm

Separate datasets were obtained from four police forces in England and Wales. All used the same definition of domestic abuse (see Chapter 9) which, while not an official crime classification, was recorded using 'flagging', in accordance with the definition. A common set of fields was established to enable the composition of an aggregate dataset. The overall dataset included 290,241 domestic abuse crimes recorded between 2009 and 2015, offered broad comparability to the overall position for domestic abuse in the whole of the country in this period (see Table 17).

Table 16 shows comparisons of each force's data in full, in relation to which several key points are worth noting. Firstly, there is a wide range of variation in the percentage of incidents that were recorded as crimes, ranging from 29% in Force B to 55% in Force A. This was not uncommon during the period analysed, when force recording of crimes was reportedly uneven (HMICFRS, 2014b). By including forces across the spectrum of rates, our dataset smooths this unevenness.

Secondly, the database shows that, while violent offences are the most frequent type of domestic abuse, they are not the only form of crime classification given the domestic abuse designation. Thirdly, the data support the ONS finding that the majority of domestic abuse victims are females, but at a different ratio; the ONS (2018) reports a ratio of around two female victims to every male, whereas these data report a ratio of 4.2 female victims to every male. Finally,

Table 16 demonstrates that the vast majority of reported domestic abuse is at the low end of the CCHI scale. These characteristics elevate the extent of generalisability of our findings because of the broad comparability with the national overview.

Table 16. Comparison of key domestic abuse statistics in Dataset 1

	Force A	Force B	Force C	Force D
Dates covered	1/4/2010–31/3/2015	1/1/2009–28/1/2015	1/4/2010–31/3/2014	1/2/2012–31/12/2015
Proportion of records that were classified as crimes	55%	29%	43%	30%
Proportion of crimes that were classified as violent	44.2%	22.4%	33.3%	28.3%
CCHI score per crime: *M* (*SD*)	34.0 (186.1)	20.9 (166.3)	41.7 (288.1)	31.1 (221.2)
Proportion of crimes that were solved	41.3%	41.4%	43.4%	41.8%
Victim age: *M* (*SD*)	29.0 (17.9)	N/A	30.7 (15.5)	33.1 (12.3)
Proportion of victims who were female	69.2%	N/A	78.4%	83.5%
Proportion of victims who were of White British ethnicity	85.7%	N/A	92.0%	66.8%
Offender age: *M* (*SD*)	32 (14.9)	N/A	33 (12.0)	33 (11.1)
Proportion of offenders who were male	54.7%	N/A	79.1%	85.1%
Proportion of offenders who were of White British ethnicity	58.1%	N/A	81.9%	61.3%

N/As indicate where the dataset provided by the force either did not provide the required variables or a high proportion of variables included missing data and summary statistics could not be produced.

Table 17. Comparisons of prevalence: Dataset 1

	England & Wales	Force A	Force B	Force C	Force D
Recorded domestic abuse per 1,000 population	17.8	14.5	17.9	11.8	18.4
Proportion of the of-age population (aged 16–59) who had been victims of domestic abuse	6.0%	6.8%	6.7%	8.1%	7.3%
Proportion of overall crime recorded by police which was domestic abuse[10]	10.8%	12%	10.7%	8.9%	10.2%

The variables obtained within these data consisted of (1) crime classification, (2) anonymised victim unique reference number, (3) anonymised offender unique reference number, (4) date that crime was reported. To these calculated fields were added, consisting of (5) Cambridge Crime Harm Index score, (6) Chronological sequence of offence for victim and (7) chronological sequence of offence for offender.

14.2.2 Dataset 2: Serial Perpetrators

Dataset 2 was obtained to address questions relating to serial offending. It comprised of crime investigations from one large police force (one of the four which supplied data for Dataset 1, but over a different timescale) identified as domestic abuse by recording officers and later validated by input clerks based on the cross-government definition of domestic abuse. The data quality of crime investigation records in the police force under examination was routinely scrutinised by audit staff, and officers were generally perceived to have a mature understanding of the definition of 'domestic abuse' following years of prioritisation of domestic abuse and investment in training to support officer awareness and knowledge.

[10] Police also record crimes in a range of other non-domestic categories – violent crime, acquisitive crime, environmental crime etc.

Table 18 shows that most crimes in the dataset did not result in an outcome which legally 'proved' that the suspect had committed the offence. When police in England and Wales complete an investigation, they are required to classify the case with an outcome code. Eight of these codes are collectively described as 'solved' outcomes, indicating the positive identification and processing of a suspect. Of these, just one (charge) results in the alleged offender progressing to court. The other seven 'solved' outcomes are known as 'out of court' outcomes, because they do not proceed to a court setting. However, the offender in these cases does receive an official criminal record. The remaining outcome codes are known as 'unsolved' cases. A suspect may be identified, but not enough evidence available to prove the crime, for example.

In this sense, the decision to expand the definition of 'perpetrators' to those with suspect status is justifiable on practical grounds. Analysis of only convicted cases would have greatly and disproportionately reduced the sample size, but it is important to emphasise that an 'unsolved' outcome is not necessarily a tacit indication of suspect innocence, just as a solved outcome is not an indication of legal guilt (as charged cases may be judged otherwise in court). As such, this research makes statements about domestic abuse perpetrators on the basis of links to crimes as suspects, which in the majority of cases are identified as part of the process of determining the status of the case as 'domestic'.

Table 18. Breakdown of domestic abuse outcomes

	Outcome type	Percentage of crimes in dataset
'Positive' outcomes (21.6%)	Charged	15.0%
	Cautioned (youth)	0.3%
	Cautioned (adult)	5.2%
	Taken into consideration	0.0%
	Offender deceased	0.0%
	Penalty notice	0.0%
	Community Resolution	1.1%
'Unsolved' outcomes (69.9%)	Not in public interest to prosecute (CPS decision)	0.5%
	Not in public interest to prosecute (police decision)	1.8%
	Named suspect below age of criminal responsibility	0.0%
	Named suspect too ill	0.5%
	Named suspect but victim deceased or too ill	0.1%
	Suspect not identified, victim cannot or will not support	0.6%
	Suspect identified, victim supports but evidential difficulties	16.6%
	Suspect identified but victim has withdrawn support	44.4%
	Time expired	0.9%
	No suspect identified, no further line of inquiry	0.6%
	Another agency progressing action	2.1%
	Further enquiries needed but not in public interest (police decision)	0.4%
Under investigation (8.5%)	Crime still under investigation at time of data extraction	8.5%

Table 19 illustrates comparability of our sample with the national trends for police forces at the time these data were recorded. These statistics indicate a high level of similarity which, while not a forensic statement of external validity, suggests that the dataset analysed here is relevant to the wider national picture.

Table 19. Dataset 2 statistical comparisons

		Year ending March 2016 (%)	Year ending March 2017 (%)
Proportion of domestic abuse recorded as a crime	Sample	35.6	44.9
	National average	40.9	45.7
Proportion of domestic abuse crimes recorded as violent	Sample	77.4	77.8
	National average	77.8	77.0
Proportion of all recorded crime classified as domestic abuse	Sample	9.7	10.2
	National average	10.8	11.0

The net result of this picture is higher confidence in the dataset as a true reflection of the actual picture of crime reported to police, even if the extent of that confidence is not quantitatively measurable.

Dataset 2 included 26,833 records taken from a period of just over two years (834 days) from October 2015 to January 2018. The dataset comprised a typical tabular format of a row of data per crime investigation, with columns relating to the different available variables, which included (1) a unique identifier for the investigation, (2) the offence classification, (3) the victim unique identifier, (4) the perpetrator unique identifier, (5) the date the crime was reported to the police, (6) the perpetrator gender, (7) the perpetrator age and (8) the perpetrator ethnicity. As with Dataset 1, several calculated fields were added, including, (9) Cambridge Crime Harm Index score, (10) victim status as serial (yes or no) and (11) perpetrator type ('serial', 'repeat' or 'single-time'). As with Dataset 1, CCHI scores were obtained by a combination of a lookup table in Excel that cross-referenced the crime classification text against a reference table, and manual data input. The latter used the central CCHI spreadsheet collated by the University of Cambridge. For perpetrator type, the category was assigned based on a combination of a count of the number of investigations linked to the

suspect unique identifier (single counts were automatically assigned to ‘single-time’, while multiple counts progressed to the next step), and then the number of unique victims. Those perpetrators with a unique victim count exceeding 1 were assigned to ‘serial’, with the remainder being assigned to ‘repeat’.

As with any police-recorded crime dataset, there are clear limitations, all of which have already been noted in earlier chapters but are worth repeating. The dataset contained only those crimes which the police were notified about, and despite the increased efforts of auditors and trainers, it is still likely that some attrition occurred between the report being made to police and a crime record being recorded. All records in this dataset were recorded on to a new computer system, implemented in full only five days prior to the extraction ‘start’ date. It is possible that some misclassifications may have occurred, but the extent of this was judged by the force to be small. The new system had very few mandatory fields, and in some cases this led to data being omitted from the records in the dataset. The final dataset was adjusted to 25,312 crime records, with 5.7% of records removed on the basis of missing or erroneous data.

14.2.3 Dataset 3: Forecasting

The objective for Dataset 3 was to test whether a machine learning technique could produce a model for forecasting future domestic abuse offending by any arrested offender, irrespective of the offence they were arrested for. Statistical modelling of this kind required a large dataset spanning multiple years to allow for ‘follow-up’ periods of time in which we were able to observe the actual offending behaviour. To this end, one large police force in England and Wales supplied a multi-year dataset of arrest records (also known as custody records).

Arrest data was used for two primary reasons. The first is that we anticipated that a high rate of ‘false positive’ or cautious errors. Such errors are generated by incorrectly forecasting that an individual will commit a domestic offence when they in fact, did not, and, in practice, may result in them receiving some form of intervention which they do not require. We believe this might reasonably be considered an infringement of the suspect’s rights and so, by selecting only arrest data, we are limiting the exposure to this forecasting instrument to only those individuals that the police have deemed eligible for arrest (i.e. they were suspected of committing a criminal offence). This of course does not mean that they were guilty of a crime, but practicality means it is impossible to obtain such information, and

conviction rates are low enough that an analysis using such a narrow dataset would almost certainly be meaningless in practice.

The second reason for selecting arrests is that they capture a wide array of persons of interest to the police, who have an identifiable checkpoint (processing through custody) with any police force where the prediction tool might be applied, but this data source captures a high proportion of 'serious' domestic abuse cases (one of the main focus points for the forecasting tool). This dataset allows us to test whether any arrestee is likely to subsequently commit a domestic offence.

In total, Dataset 3 comprised of 73,380 arrest records. These records were case events, not individual records, so it was possible for an offender to be included multiple times in the data. Analysing records this way has more potential for practical relevance in that an agency may wish to forecast and re-forecast an individual upon each time they enter custody. At each new arrest their criminal history may have changed and therefore the probability of future offending may change with the number of times an individual is arrested.

The dataset included arrest records from 1999 to 2017. In its raw form the data consisted of multiple tables of data, linked together by shared unique reference numbers (URNs). These tables were assembled into one training dataset with variables for (1) arrestee URN, (2) crime classification (for which they were arrested), (3) date of arrest, (4) gender of the arrestee and (5) arrestee date of birth. Calculated variables (those created by calculations based on supplied variables – e.g. to create a variable for 'presenting arrest is for a domestic crime', code was written to apply a '1' were the domestic keyword was present, and otherwise a '0'), were added to these to formulate the required set of predictor variables (those used by the statistical model to forecast the outcome) and one calculated variable was added to represent the outcome the model would attempt to forecast. This variable ('outcome') was coded to either 'DV0', 'DV1' or 'DV2'. Arrestees with no domestic arrest within two years were coded as 'DV0'. Those with a less-serious domestic arrest were coded as 'DV1' and those with a subsequent arrest for a serious domestic offence were coded as 'DV2'.

The full set of 35 predictor variables were as follows:

- **Age at custody event:** the age in years of the suspect on the day of the presenting custody event in each case (note that each offender can appear in multiple cases, but their age in each case may vary depending on when they are arrested;

moreover, each subsequent crime presents a new two-year follow-up from the new date of arrest)

- **Gender:** female or male
- **Number of prior crimes:** the number of crimes attributed to the offender prior to the presenting custody event case
- **Number of presenting crimes:** the number of crimes attributed to the offender in the presenting custody event
- **Presenting crime charge is for violence:** yes/no
- **Presenting crime charge is for a property crime[11]:** yes/no
- **First age – any offence:** the age of the suspect in years on the recorded date of their first crime
- **First age – serious offence:** the age of the suspect in years on the recorded date of their first serious crime, regardless of when it occurred (serious as defined in outcomes; see footnote 1)
- **First age – weapons offence:** the age of the suspect in years on the recorded date of their first crime involving a weapon
- **First age – drugs offence:** the age of the suspect in years on the recorded date of their first drug crime (possession or supply)
- **First age – property offence:** the age of the suspect in years on the recorded date of their first property crime
- **Years since last offence:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed offence to the date of the presenting custody event
- **Number of prior murders:** the number of homicides and attempted homicides attributed to the offender prior to the presenting custody event
- **Number of prior serious crimes[12]:** the number of serious crimes attributed to the offender prior to the presenting custody event
- **Years since last serious crime:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed serious offence to the date of the presenting custody event

[11] Property crimes include burglary, theft and criminal damage offences
[12] Serious crimes were defined as higher than 545 days on the Cambridge Crime Harm Index scale.

- **Number of prior violent crimes:** the number of violent crimes attributed to the offender prior to the presenting custody event
- **Years since last violent crime:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed violent offence to the date of the presenting custody event
- **Number of prior sexual crimes:** the number of sexual crimes attributed to the offender prior to the presenting custody event
- **Years since last sexual crime:** the number of years (rounded to whole number) from the recorded date of the offender's last attributed sexual offence to the date of the presenting custody event
- **Number of prior weapons crimes:** the number of weapons crimes attributed to the offender prior to the presenting custody event (based on offence category, not keyword)
- **Years since last weapons crime:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed serious offence to the date of the presenting custody event
- **Number of prior firearms crimes:** the number of firearms crimes attributed to the offender prior to the presenting custody event
- **Years since last firearms crime charge:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed firearms offence to the date of the presenting custody event
- **Number of prior drug possession crimes:** the number of drug possession crimes attributed to the offender prior to the presenting custody event
- **Years since last drug possession:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed drug possession offence to the date of the presenting custody event
- **Number of prior drug supply crimes:** the number of drug supply crimes attributed to the offender prior to the presenting custody event
- **Years since last drug supply crime:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed drug supply offence to the date of the presenting custody event
- **Number of prior property crimes:** the number of property crimes attributed to the offender prior to the presenting custody event

- **Years since last property crime charge:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed property offence to the date of the presenting custody event
- **Number of prior custody events:** the number of custody events attributed to the offender prior to the presenting custody event
- **Years since last custody event:** the number of years (rounded to a whole number) from the date of the offender's last custody event to the date of the presenting custody event
- **Number of prior serious crimes:** the number of serious crimes attributed to the offender prior to the presenting custody event
- **Years since last serious crime:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed serious offence to the date of the presenting custody event
- **Number of prior domestic crimes:** the number of domestic crimes attributed to the offender prior to the presenting custody event
- **Years since last domestic crime:** the number of years (rounded to a whole number) from the recorded date of the offender's last attributed domestic offence to the date of the presenting custody event
- **Violent warning marker:** presence of a warning marker in the same year[13] as the presenting custody event (yes/no)
- **Drugs warning marker:** presence of a warning marker in the same year as the presenting custody event (yes/no)
- **Weapons warning marker:** presence of a warning marker in the same year as the presenting custody event (yes/no)
- **Suicide warning marker:** presence of a warning marker in the same year as the presenting custody event (yes/no)

14.3 Procedure: Repeat abuse

Research questions 1 and 3 (relating to the prevalence of repeat victims and offenders in Dataset 1), required calculations of the percentage of victims and, separately, offenders for whom the total count of crime records was greater than one. This process was undertaken

[13] For warning markers the presence of a marker in the same year as the presenting event was selected to account for contemporaneity. For example, an individual who has been suicidal ten years ago may not longer be so. Our parameter might reasonably be extended to a longer period of time.

using a pivot table in Excel, which was also used to provide a breakdown of the frequency for each total crime count (i.e. how many offenders had two crimes, three crimes and so on).

Questions 2 and 4 required the calculation of conditional probability. This was calculated in the normal way (P(A|B) – i.e. the probability (P) of an event (A) occurring given that another event (B) has already occurred) by establishing a frequency table for total crime counts (as described in the previous paragraph) and then dividing the summed frequencies for each respective count. For example: to calculate the conditional probability of a victim reporting their second domestic crime subsequently reporting a third, we calculated the total frequency of victims reporting 3, 4, 5 …*n* crimes, and then divided this by the total frequency of victims reporting 2, 3, 4, 5 …*n* crimes.

14.4 Procedure: Serial abuse

Questions 5 and 6 (the prevalence of serial victims and offenders) require descriptive statistics in the same fashion as Questions 1 and 3. To this end, datasets 1 and 2 were both analysed (both offering the opportunity to calculate the prevalence of serial victims and offenders), by use of the victim/offender type variable and pivot tables to calculate the proportion of all victims/offenders with a distinct offender/victim count greater than one (e.g. the serial rate for victims was the total number of victims linked to >1 offender, divided by the total number of victims).

Question 7 (relating to demographic differences between perpetrator types) draws only on Dataset 2 age, ethnicity and gender variables which are then compared between cohorts using independent sample t-tests and chi squared tests to determine if the differences were statistically significant. While Dataset 1 contained unique identifiers that enabled us to calculate the prevalence of serial perpetration, it did not contain enough demographic information for us to analyse serial offenders in detail, hence obtaining these data in full for dataset 2 and using only this source to address this question.

For Question 8 (the profile of domestic abuse crimes committed by ‘serial’ perpetrators compared to ‘repeat’ and ‘single-time’ offenders), the analysis required a sub-group analysis of crime classification counts by differing offender types. Crime classifications were grouped into categories based on Home Office reporting rules. These results were also analysed using independent sample t-tests to establish any statistical significance in the differences. This approach was mirrored for Questions 9 and 10, using CCHI scores for non-domestic crimes.

For question 9, we calculated to the 'power few' for the overall dataset (the proportion of offenders which contribute a cumulative 80% of total CCHI) and then examined the relative contributions to that group by each offender category.

14.5 Procedure: Escalation

Given that 'escalation' in the domestic abuse context, implies a continuing pattern over time, our analytical procedure to identify it was undertaken on repeat and serial cases (for both victims and offenders) with no fewer than five total crime records in Dataset 1. Five crimes was the threshold chosen based on the sample size this provided (see Table 20 for the respective sample sizes at each total count increment for victims and offenders), together with the scope that five separate incidents offer to establish a pattern of harm. Our logic here was as follows: for us to be able to detect escalation we needed multiple data points, and enough of them to enable some form of pattern to be established.

Table 20. Sample sizes for each category of chronological crime analysed for escalation: victims

Crime	1	2	3	4	5	6	7	8	9	10
Victims (n)	8,704	8,704	8,704	8,704	8,704	5,867	4,160	3,007	2,230	1,738
Offenders (n)	9,337	9,337	9,337	9,337	9,337	5,723	3,960	2,683	1,912	1,348

The analysis tracked the mean CCHI scores of all eligible victims (question 11) and offenders (question 12) up to their tenth crime. After ten crimes the sample sizes became too small to be meaningful. To conduct the analysis, eligible victims and offenders were collated, and tables created for the mean CCHI score for the first to tenth offences. These calculated data were then analysed with the one-way ANOVA test to determine whether there was any statistically significant difference present. The post-hoc test Tukey's Honestly Significant Difference (HSD) was then applied to establish between which crimes in the sequence that the significant difference existed.

14.6 Procedure: Concentration of harm

Questions 13-16 were addressed using interpretations of pivot tables from Dataset 1. We did not use datasets 2 or 3 for this purpose because of smaller size and scope, respectively. Dataset 2 was just one force and two years of crimes. Dataset 3 had a number of years but was only arrests.

Both victims and offenders were reorganised into tables reflecting the descending order of total CCHI score for individuals. A cumulative total CCHI was then recorded, along with a cumulative for the number of individuals. The 'power few' threshold was applied at 80% cumulative total CCHI because of the traditional use of this figure in analysing 'J curves' or Pareto curves and the common use in many fields of the '80-20 rule of thumb' which attributes 80% of an measurement to 20% of total units (Chen, Chong and Tong, 1993; Cohen and Mandrack, 2000; Iqbal and Rizwan, 2000; Sherman, 2007; Weisburd, Groff and Yang, 2012)

14.7 Procedure: Forecasting

Up to this point, the analytical procedure outlined was made up of primarily descriptive statistics. But to build a statistical model capable of accurately forecasting domestic abuse among a population of arrestees for any type of crime (Question 19), we required a more complex procedure. As outlined in Chapter 4, our chosen method is a random forest algorithm (Breiman, 2001). The following subsections outline the main principles of the random forest procedure and give the specifics of the process as it was applied. Additional technical information is documented in Appendix I, while Chapter 19 focuses mainly on the principal results around model accuracy.

14.7.1 How random forest algorithms work

The random forest algorithm is a supervised machine learning procedure, meaning that it leverages the computational power of modern computers to process millions of calculations and make informed decisions, but with overall control by a human (Breiman, 2001; Jordan and Mitchell, 2015). A random forest process is used for classification or regression calculations, but we were interested only in the classification variety because the forecasts in question sought to 'classify' the potential outcome in respect of future domestic abuse arrests into categories.

As discussed in Chapter 12, the random forest algorithm is a form of 'decision tree' modelling (Liaw and Wiener, 2002). An application of the procedure works by constructing multiple iterations of 'decision trees' (which when combined create the 'forest'). A 'decision tree' is a common decision-making framework used in forecasting and the visualisation of algorithms. A tree is formed of two components: (1) decisions and (2) nodes. In the context of this research a 'decision' relates to a specific variable, such as the gender of the arrestee. Nodes are created by the differing responses to a decision – so in this example the decision would create two nodes: (1) female and (2) male. Each node has a 'weight' reflected by the proportion of the node which is attributed to the overall outcome. In this example, the outcome we seek is 'no domestic abuse' or 'some domestic abuse' within a specified timeframe. Hypothetically, the decision and nodes may look like Figure 3.

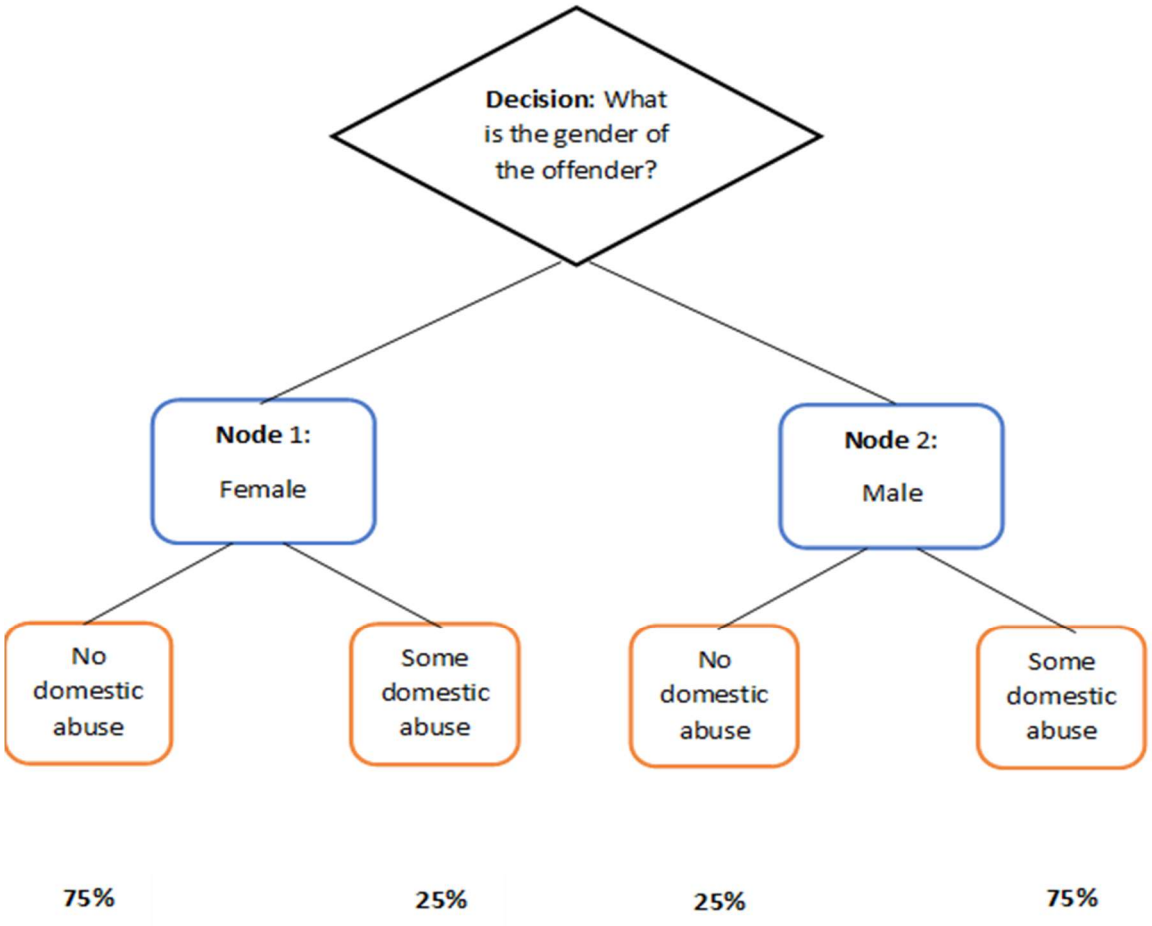

Figure 3. Example of a basic decision tree

In such a simplistic example, if the arrestee were a male, the algorithm would forecast domestic abuse, given that the proportion of male arrestees with future domestic abuse in its training dataset was 75% and for females it was just 25%. In this sense the outcome with the

greatest probability determines the forecasting decision – and this is the central 'decision' principle of a decision tree model. Conversely were the model applied to a female arrestee, the forecast would be for no future domestic abuse. Based on the training data we would expect these forecasts to be correct around three quarters of the time. The remaining 25%, that is, the 'errors' may be classified in two forms: (1) forecasts of domestic abuse which proved incorrect – we refer to these as 'false positives' or 'cautious errors' and (2) forecasts of no domestic abuse which proved incorrect – these are also called 'false negatives' or 'dangerous errors' (see Banerjee, Chitnis, Jadhav, Bhawalkar and Chaudhury, 2009)

In practice the application of the algorithm is far more complex, but all of these principles remain. A random forest algorithm works by constructing multiple decision trees (the researcher can specify the number), and by selecting multiple 'predictor variables' on which to form the decisions in that tree (Breiman, 2001). The researcher can specify how many of these are selected in each tree, but the selection of which variables are used is random (hence the 'random' in random forest). Adding even just one additional decision point substantially increases the number of possible decision pathways, as depicted in the hypothetical decision tree shown in Figure 4.

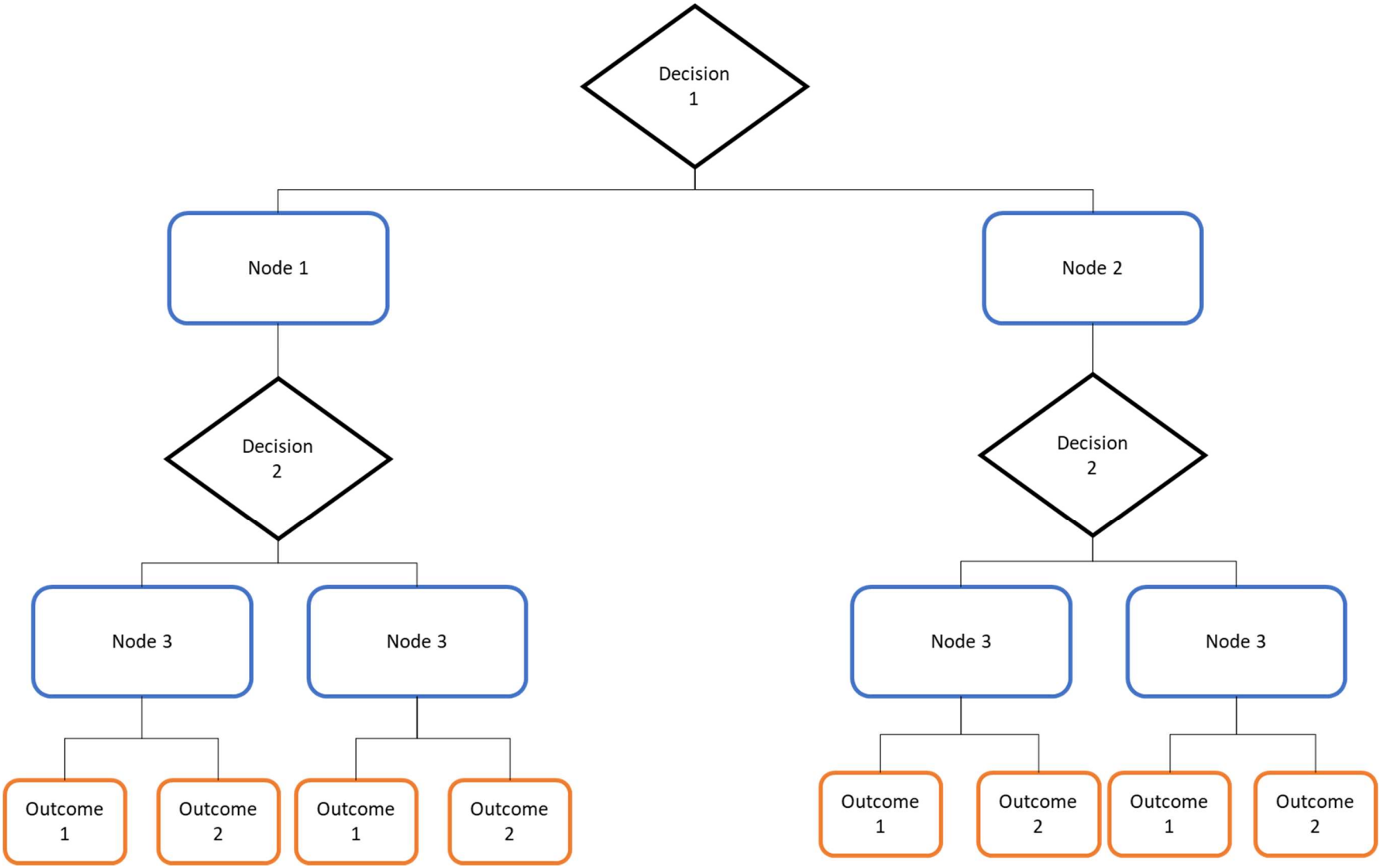

Figure 4. Example of a decision tree with two decision points

By adding a single extra decision variable, the 'decision tree' has twice as many outcome pathways (eight compared to four) but the forecasting principle remains the same: the pathway with the greatest probability associated to an outcome (based on the training dataset) is selected as the forecast. For instance, consider that our first decision point variable is gender, and our second is whether the index arrest is for a domestic crime (either a 'yes' or a 'no'). Each case in the training data is assessed on these two variables and follows a branch accordingly. A case in which the offender is female, and the presenting offence is not domestic would pass on the leftmost branch. If the offender was female and the presenting offence was a domestic then it follows the next branch to the right, and so on. Of course, in this simplified example, each decision has only two possible answers (nodes). In practice, a variable may have more. Each record in the training data is computed in this way, and the proportions of each branch outcome are used by the algorithm in the forecasting process.

Naturally then, the algorithm is only as good as the data on which it is trained. The random forest procedure works by constructing hundreds of these trees and 'learning' which predictor variables offer the best forecasting accuracy as it proceeds. With each new tree constructed the forecasting accuracy should improve (up to a finite point, because no model is ever likely to be perfect – see Kulkarni and Sinha, 2012 for more detail).

The random forest procedure is often favoured by researchers because of its method of self-validation – the method by which it 'learns'. It is typical for statistical forecasting models to suffer from a phenomenon known as 'overfitting', which describes artificially high forecasting accuracy because of either the uniqueness of the training dataset in comparison to the real world, or by including too many parameters into the statistical calculations than can be justified (Horning, 2013; Liaw and Wiener, 2002). Random forest algorithms approach this by partitioning a segment of its training data. With the construction of each new tree, the algorithm randomly selects around two-thirds of its training records on which to base the tree. The other third is then used to test the accuracy of the tree when applied to records not involved in the training process (Breiman, 2001). This testing sample is referred to as 'out-of-bag' (known as OOB) and the OOB error rate - the proportion of OOB records for which the decision tree did not correctly forecast the outcome - is tracked by the algorithm to refine the accuracy of the subsequent trees it builds. As it develops more trees, the algorithm learns which combinations of variables yield the best results and refines its selection of variables included in trees accordingly. In this way, our forecasting model used Dataset 3 as its training dataset and evaluated its own performance as part of its construction.

Finally, for overall evaluation, each record in Dataset 3 was 'dropped down' each of the decision trees in the forest to determine the forecasted outcome (see *Outcome*). The forecast of each tree was considered as an equal 'vote' and the outcome with the overall majority of votes was the resulting forecast of the model.

14.7.2 Model parameters

Model 'tuning' was undertaken in the statistical package, 'R', to determine the number of trees and number of decisions. 'Tuning' involves repeatedly re-running models with incremental changes to parameters in a search for optimal forecasting accuracy. The results of these processes are included in *Appendix A: Technical Information Relating to Random Forest Modelling*. The number of trees was set to 501 and the number of decisions in each tree set to four. The appendix explains how these parameters were selected by examining the overall average out of bag error rates for each outcome class, for each incremental tree. Once these error rates stabilised we determined it was appropriate to cap the number of trees.

The following subsections outline the different aspects of the model developed to address Question 19. Our general approach takes its cue from two sources: (1) Gottfredson and Moriarty (2006), which set out 10 common methodological pitfalls often overlooked in the construction of forecasting instruments; and (2) Barnes and Hyatt (2012) which, as described in Chapter 12, is the most detailed presentation of a random forest procedure in a criminal justice setting to date. The next paragraphs summarise the approach taken in the construction and assessment of the model, drawing on these sources.

14.7.2.1 Instrument selection

The random forest ensemble classification tree method was used to construct our forecasting model based on the reasons outlined in *12.6*. The open source statistical package 'R' was used to apply the method, using the specialist *randomForest* package.

14.7.2.2 Definition of unit of prediction and follow up period

For the sake of clarity, the unit of prediction was defined as an arrested offender. The follow up period for the forecast was 24 months. This period was chosen as it replicates previous studies (Berk and Hyatt, 2012; Berk, Barnes and Sorenson, 2012) and represents a practical working period which is neither so short as to marginalise the outcomes nor so long as to prohibit meaningful analysis of a follow-up period.

The period 2012 to 2015 was used for two reasons. Firstly, domestic abuse 'flagging' on arrest and crime records was much less frequent (inaccurately so, we believe) before 2012. Secondly, by choosing 2015 as the final cut off, we allowed for a clear 24 months of follow up in which to observe the outcomes (up to the end of 2017 – the time at which the dataset was obtained). All years prior to 2012 were included for the calculation of predictor variables (i.e. how many prior violent arrests an offender had included records back to 1999).

14.7.2.3 Outcome variable

The statistical model was built with three categories in the outcome variable to be forecast: (1) no domestic abuse arrest within two years (DV0), (2) an arrest within two years for a less-serious domestic crime, or (3) an arrest within two years for a serious domestic crime[14]. The definition of serious encompassed violence with serious injury or homicide (including attempts) and sexual offences, broadly in keeping with the definitions used by Barnes and Hyatt (2012), Berk, Sorenson and Barnes (2012) and Urwin (2016). This definition could easily be amended to suit the requirements of any organisation, but as this model is an illustrative test of this concept, we selected a definition that will allow broad comparability with predecessors and is in keeping with other findings on concentration of harm (see Chapter 18).

14.7.2.4 Predictors

A key advantage of the random forest method is that it places no limit on the number of predictor variables that can be included in modelling, nor do they require prior causal association with the outcome variable. As Barnes and Hyatt (2012) explained, predictor variables with less influence are used less in the final classification trees selected by the algorithm, but even the less influential variables can marginally boost accuracy with no real cost.

While there is technically no limit to the range of predictor variables that can be chosen, we have opted for a suite of variables that are composed solely from police data and mostly relate to characteristics of an offender's criminal history (see 14.2.3 for the full list of 35 predictor variables chosen). We have tried to place emphasis on data that any police agency should store.

[14] Serious was defined as all crime classifications with a CCHI value of 545 or above

We have deliberately omitted predictors which may be considered to introduce overt bias such as ethnicity, nationality, disability or socioeconomic classification. As we have already touched on, there is a looming controversy over the police use of algorithms based primarily on concerns of unfairness and lack of transparency (Burrell, 2016; Liberty, 2019; Oswald et al., 2018; Pasquale, 2015). These are serious issues, not least because public support for policing activities is intrinsic to their success and the population is lawfully entitled to respect for human rights (Neyroud, 2012). As an emerging technique, closely related to the development of new technology, the application on machine learning algorithms is particularly susceptible to the risk of damaging public confidence described by Neyroud and Disley (2008).

As established by previous authors of random forest research (see Berk, Sorenson and Barnes, 2012, for example), it can be useful for the commissioners and users of a forecasting tool to understand which predictor variables make the greatest contribution to forecast accuracy. This analysis must be handled with some care, because random forests cannot indicate causal links between predictor and outcome variables. Yet there is value to be derived from examining variable influence in building a compelling case for the effectiveness of a model, hence the inclusion of Question 20 – which predictor variables contribute the most to forecasting accuracy?

To address this question, 'variable importance plots' and 'partial response plots' were produced – both being native functions of the R package '*randomForests*'. Importance plots show the average decline in model accuracy and the mean decline in Gini index (Breiman, 2001; Liaw and Wiener, 2002). The mean decline in accuracy is calculated by re-running calculations without each variable in turn and storing the performance of each iteration. The Gini index refers to 'node purity', which is a measure of how related a variable is to one particular outcome class. Recall that each split in a single decision tree is a node, from which the algorithm (randomly) selects a predictor variable. The algorithm improves its own performance by selecting variables with the lowest Gini indices. A low index indicates that a variable favours a particular outcome class; for example, if all female arrestees went on to commit no domestic abuse, then the Gini index would be 0.

The model logic is summarised in Chapter 19, with full details including samples of plots, in *Appendix A: Technical Information Relating to Random Forest Modelling.*

14.7.2.5 Relative costs of errors

In the calculation of a random forest model, the researcher has the ability to offset the 'relative costs' of forecasting errors, thus introducing some measure of deliberate bias into calculations against the 'costlier' outcome of an error (see Berk, 2012). In this case, that outcome would be the occurrence of an arrest for serious domestic abuse (DV2) when none (DV0) was forecast (as opposed to the occurrence of no domestic abuse when serious abuse was forecast, which would be a 'cautious' error rather than a 'dangerous' on the basis that resources were assigned when they need not have been). Given the emphasis on forecasting of serious domestic abuse, we opted to set equal sample limits for each outcome classification. In practice this means that the random forest algorithm selected 400 records of each outcome type on which to train each tree. This allocation favours the serious outcome, which is subsequently proportionally oversampled relative to the other two outcomes. The expected result would be a greater rate of 'cautious' errors and a lower rate of 'dangerous' errors because the training sample was deliberately disproportionately comprised of the most serious crimes, skewing forecasting accuracy in favour of that outcome.

14.7.2.6 Presentation of findings

In the presentation of the results of the modelling, we attempt to cover many of the bases recommended by previous studies. Accuracy is presented from various perspectives: overall accuracy, the proportion of each outcome level that is accurately forecast, and the proportion of each forecast that is correct. The results are presented in the form of the traditional 'confusion matrix' commonly used in forecasting papers (see Berk, 2012) and supporting tables, as used in Barnes and Hyatt (2012) and Urwin (2016). This 'matrix' is essentially a cross-tabulation of the forecasted outcomes for each record in the training dataset, referenced against the actual outcomes. It also summaries the accuracy of each type of forecast (i.e. how many were correct) and the efficiency of the model (i.e. how many of each type of actual outcome were correctly forecast).

14.7.2.7 Baseline considerations

To place the accuracy of the forecasting model in context we need to understand the baseline occurrence of the outcomes in question. Baseline occurrences are the extent to which each outcome actually occurs in the training data. In our case we are interested to know what the actual rates of no future arrest for a domestic crime, an arrest for a less serious domestic crime and an arrest for a serious domestic crime are. These will set out expectations for the effectiveness of the model. If 98% of cases actually had no future arrest, then a forecast rate

of no future arrest at 95% would not seem unreasonable, but we would not know this without first establishing the baseline rate. These are captured by research questions 17 and 18 and are answered by descriptive statistics relating to prevalence.

14.8 Summary

Three large datasets obtained from multiple police forces across England and Wales were accessed in order to address the research questions set out in Chapter 13. The analytical procedures we have described for the assessment of repeat and serial abuse, escalation and the concentration of harm are predominantly descriptive – the power of our analysis is drawn from the scale and nature of our datasets in these cases, rather than the nature of our statistical procedures. We also described the application of difference tests in the comparison of different groups, notably with respect to serial abuse, for which we seek to compare characteristics of serial domestic abuse offenders with those of single-time and repeat offenders.

We have also described, in some detail, the more complex procedure of developing a forecasting model using the random forest algorithm. This procedure constructs hundreds of decision trees, each tested against a randomly drawn independent sample, in order to refine an overall statistical model. Our model assesses arrest cases – for *any* type of criminality – and makes a forecast of future arrest for domestic abuse – either no arrest, an arrest for a less-serious domestic crime, or as is our specific focus, an arrest for a serious domestic crime. The procedure in this chapter explains how the results are presented with particular reference to baseline rates (the rates at which each of these outcomes *actually* occur) and the different forms of forecasting error. We have also discussed how the influence of individual predictor variables is assessed and presented and contextualised why it is important to consider these factors for purposes of demonstrating legitimacy, even though the statistical model does not concern itself with cause and effect.

The next chapters present the results of these procedures and attempt to do so in an accessible way. Additional technical information about the random forest model is included in *Appendix I.*

15 Repeat Abuse Findings

15.1 Chapter roadmap

This chapter presents the findings of the analysis of Dataset 1 in relation to research questions 1-4. It first considers the prevalence and probability of repeat victimisation, then offenders. The results show that around three quarters of domestic abuse reported to police involves victims and offenders just once. However, of the quarter of victims and offenders with multiple reports, there is a pattern of chronic abuse. As soon as a victim or offender is linked to a second domestic abuse crime, the probability of a third crime becomes more likely than not. The probability of additional crimes then increases with each additional report – the more a victim calls, the more they will call again.

15.2 Prevalence of repeat domestic abuse among victims

Dataset 1 contained 170,391 unique victims in total and the vast majority of them were linked to just one reported crime, as shown in Figure 5 and Figure 6.

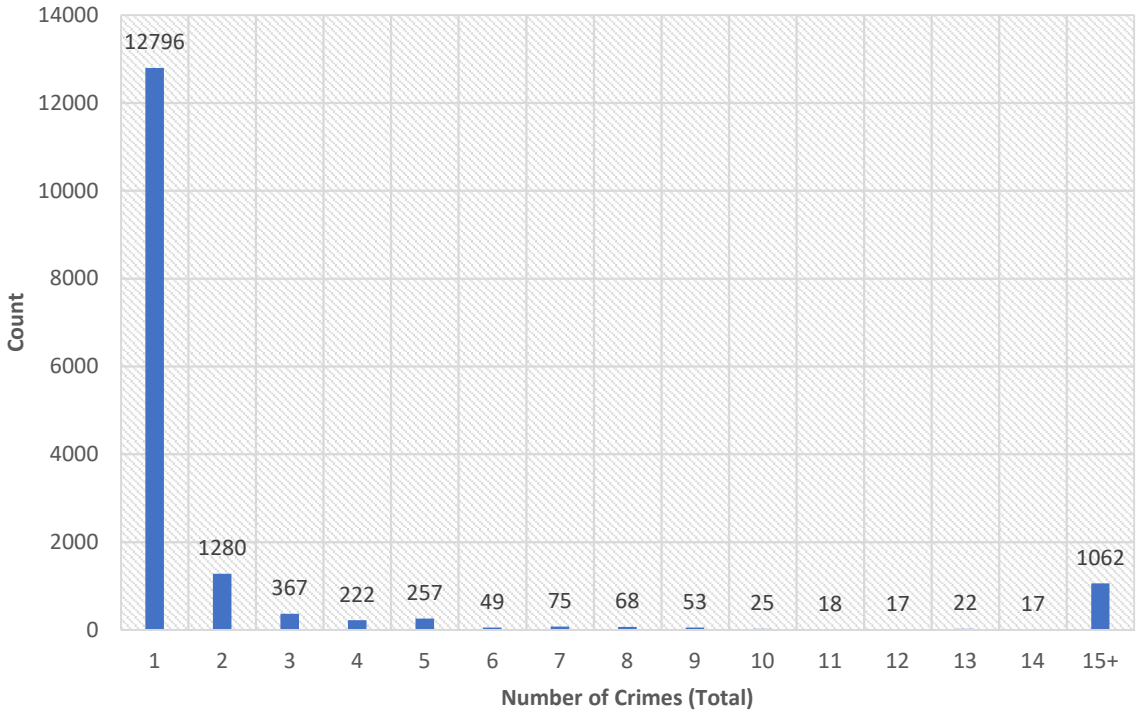

Figure 5. Number of unique victims by number of crimes recorded

Figure 6 shows that almost three quarters of these victims reported just once in a multi-year period; 25% of victims were repeats, and just 5% of victims reported three or more

domestic abuse events. Although some cases were extremely chronic – reporting in double figures, such victims were rare (0.3%).

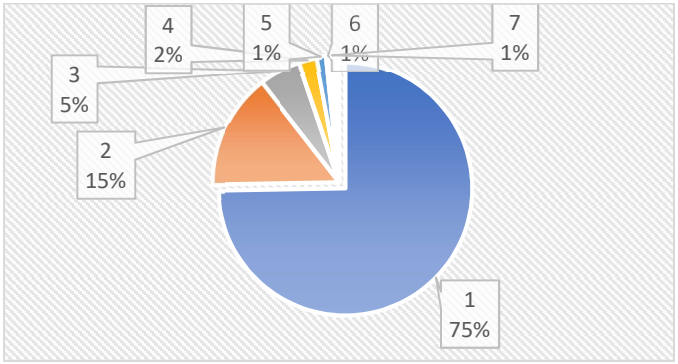

Figure 6. Percentage of unique victims by number of crimes recorded

15.3 Conditional probability of further crimes for victims[15]

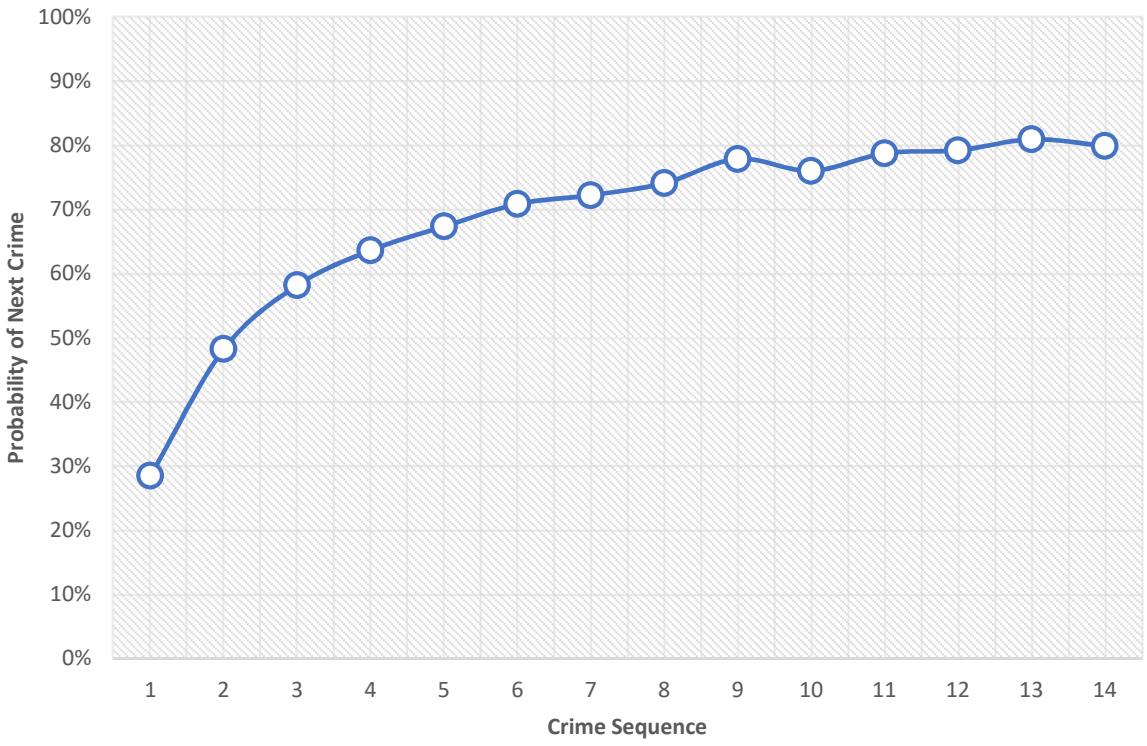

Figure 7. Conditional probability of victims being attributed to another crime

[15] Note that the scale runs to the 14th crime in sequence. Though 535 victims had greater totals than this, the sample size for each individual count was lower than 100 and so the data are omitted.

As the repeat rates for victims was 25%, so was the probability that a victim presented a second time. After this the probability rose steadily, to around a 50% chance of a third event after the second, and an over 80% chance of additional calls after the twelfth. The broad pattern, which can be observed in Figure 7, is one of increasing probability of another domestic abuse event with each additional event that occurs albeit with plateaus between 70 and 80%. The data suggest however, that the general probability of further domestic abuse increases with additional crimes.

15.4 Prevalence of repeat domestic abuse among offenders

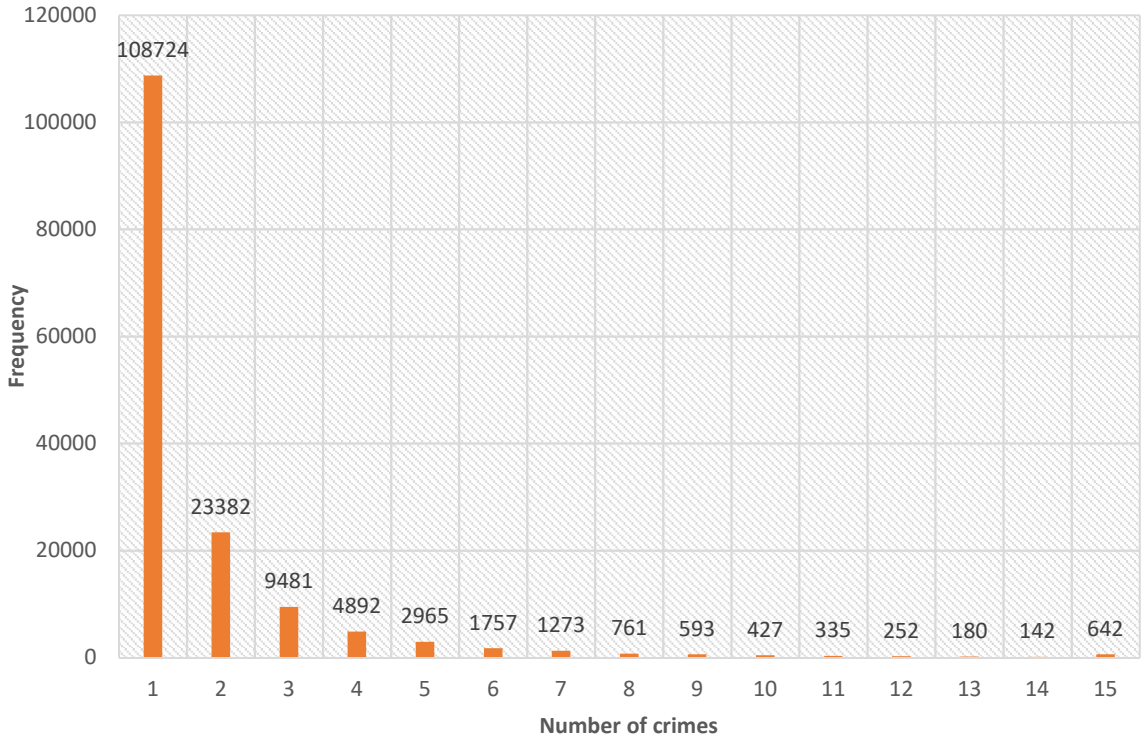

Figure 8. Number of unique offenders by number of crimes recorded

Figure 8 shows the distribution of frequencies for the 155,590 unique offenders in Dataset 1. The lower overall number in itself suggests a greater tendency toward repeat patterns among offenders, and as Figure 9 shows, the proportion of any repeat offenders was 5% higher. Broadly though, the prevalence of repeat abuse for offenders was similar to victims, with the exception of the proportion of individuals with 15 or more events attributed, which was slightly higher at 0.4% of offenders compared to 0.3% of victims.

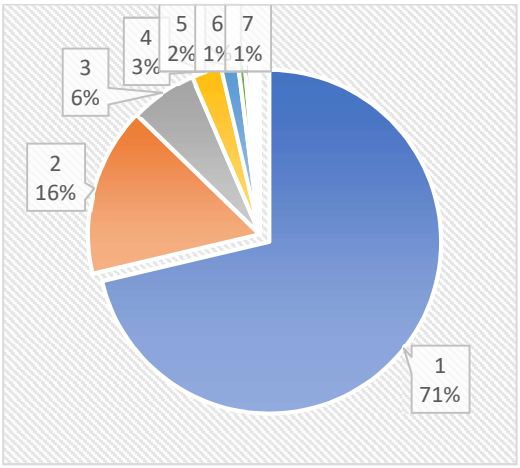

Figure 9. Percentage of unique offenders by number of crimes recorded

15.5 Conditional probability of further crimes for offenders[16]

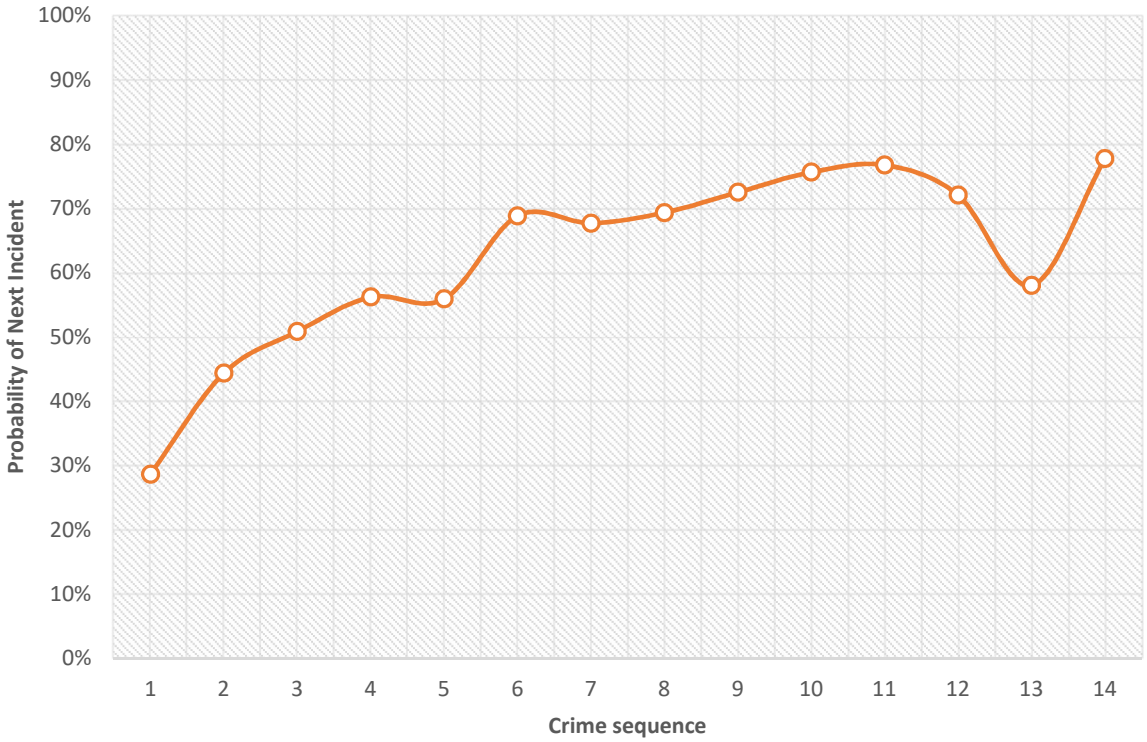

Figure 10. Conditional probability of offender being attributed to another crime

As with victims (Figure 7), the probability of offenders being linked to further domestic abuse crimes generally increased with each additional crime. Although there was a slight decline between offences 4 and 5, there is no obvious reason, and the pattern is corrected by

[16] Although 642 offenders had 15 or more total crimes, the sample sizes for each were lower than 100 offenders, so these cases are omitted from the chart.

offence 6. The variation at offence 13 is attributed to lower sample size and overall the probability of further abuse was higher from offence 6 onwards for offenders than victims.

15.6 Summary

Simple though these statistics may be, the findings are clear – most domestic abuse recorded by these four police forces did not involve a repeat offender or victim. However, where repeat cases did exist, the probability of additional domestic crimes tended to increase with each additional report up to an approximate ceiling of 80% probability. The implications for these findings are explored further in Chapter 12.

16 Serial Abuse Findings

16.1 Chapter roadmap

In this chapter we cover the findings of analysis in relation to research questions 5 to 10. We begin with a subsection on the prevalence and profile of serial abuse and its perpetrators, which draws on both datasets 1 and 2. Dataset 1 comprises of more total records, and covers longer period of time contains no information about non-domestic offending by domestic offenders. Dataset 2 was obtained specifically to examine serial perpetrator characteristics for non-domestic abuse crimes as well domestic crimes, so this is the primary focus of the chapter. However, as both datasets provide the data required to identify prevalence, both have been used to address these questions.

Dataset 2 is analysed to describe the distribution of offenders into three categories, 'single-time, repeat with one victim and repeat with multiple victims (the latter are what we refer to as serial perpetrators). The chapter presents the differences in basic demographic characteristics of these categories and the differences in the groups for both domestic and non-domestic abuse crime harm patterns. The findings show prevalence of between 10% and 15% and that serial perpetrators have distinctly different offending trends compared to repeat or single-time perpetrators. Serial perpetrators contribute to the most harmful crimes but no more so than repeat offenders though it is notable that serial perpetrators commit more crime and more crime harm than the other groups when it comes to non-domestic crimes. This opens up a further branch of sub-classifications in which we cross refer our three offender categories by three categories related to their non-domestic offending patterns. This perspective shows that repeat and serial generalist offenders are the most harmful domestic abusers.

16.2 Prevalence and profile

16.2.1 From Dataset 1

A total of 45,088 individuals were identified within the database as having serial[17] involvement in domestic abuse. 21,391 victims, representing 13% of all victims and 44% of all repeat victims, had multiple designated offenders. The rate of serial offending was higher:

[17] Serial involvement is defined as multiple crimes with differing other parties. At a minimum a victim or offender would feature in two crimes, each with a different victim or offender, respectively.

23,697 offenders were identified as serial, which represented 15% of all offenders and 50% of all repeat offenders.

16.2.2 From Dataset 2

As Figure 11 shows, the extent of repeat offending in the dataset was very similar to that of the pattern observed in Dataset 1. However, whereas half of the offenders in that analysis offended against more than one victim, only 40% did in these data.

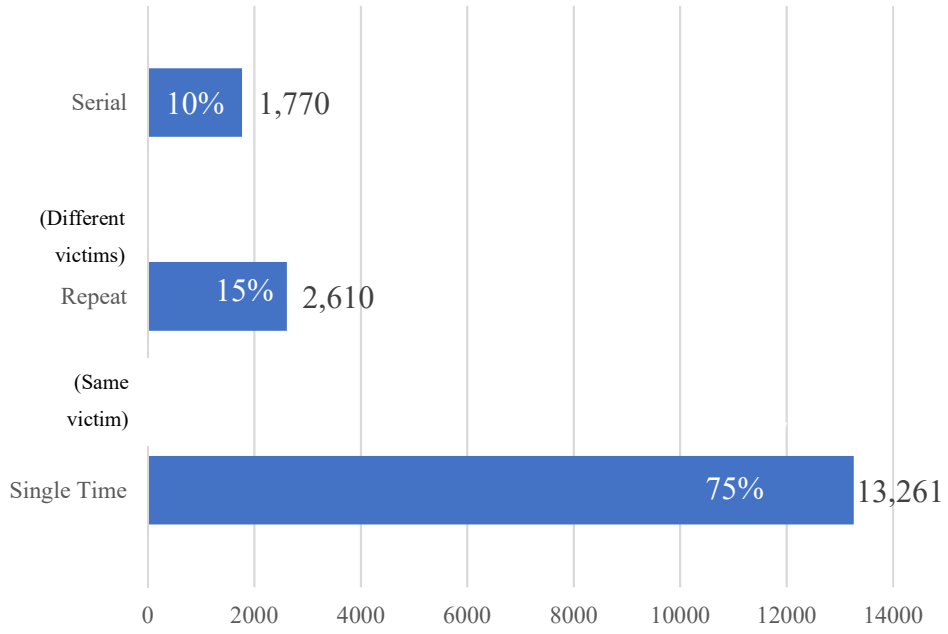

Figure 11. Offender cohort frequency

The difference from the Dataset 1 analysis might be explained by the shorter timescale of these data (they cover approximately half of the period of Dataset 1). It is logical that an offender may accumulate additional victims in longer follow up periods. However, this result does add to a repeated 'clustering' in research studies of prevalence of serial perpetrators at around 10–15% of all domestic abuse offenders (Bland and Ariel, 2015; Hester and Westmarland, 2006; Robinson, 2017).

Table 21. Selected demographic characteristics of perpetrator cohorts

Cohort	Mean age at time of offence	% male	% non-white British
Single	34.9***	75%	18%
Repeat	34.5	82%	13%
Serial	33.8	83%	18%
Statistical significance of difference compared to serial perpetrators ** p< .05, ** p<.01, *** p<.00*			

Table 21 indicates some minor differences between serial and non-serial perpetrators. It shows the mean age of serial offenders (n = 1,770) to be a year below that of single-time offenders (n = 13,261) to a statistically significant level (t *(15,020)* = 3.36, p = 0.0008*)* indicating that serial perpetrators tend to be younger than single-time perpetrators, but not younger than repeat offenders. Although females were more frequently single-time perpetrators, there was no significant difference in the proportions detected by a test for proportions. The rate of non-white British perpetrators was the same in the serial and single cohorts, but higher in the serial cohort than the repeat cohort by a ratio of 1.38:1, but again with no significance found by a t-test for proportions.

16.3 Types of Abuse

Table 22. Breakdown of makeup of domestic abuse crime types by cohort

Cohort	Violence without injury	Violence with injury	Criminal damage	Rape	Other	Total
CCHI (M)	*8.4*	*104.7*	*3.2*	*1,848.3*	*56.9*	
(SD)	*52.2*	*339.1*	*17.8*	*158.8*	*177.0*	
Single	52.0%***	27.7%	7.1%***	4.2%***	9.0%	100%
Repeat	51.8%**	28.2%	8.2%	4.1%***	7.7%	100%
Serial	49.4%	27.0%	11.1%	1.8%	10.7%	100%
Statistical significance in difference compared to serial perpetrators *** *p*< .05, ** *p*<.01, *** *p*<.00**						

Table 22 shows the variance in the composition of crime types between the three offending cohorts. For reference, the first row of the table displays the average CCHI value in the corresponding classification group, indicating the relative levels of harm. The second row displays the standard deviation, indicating the level of variance within the individual crime types which compose the classification group. The 'other' category represents all crime classifications not included within the four classifications to the left thereof (such as fraud, theft or burglary). The bottom three rows of Table 22 show the proportion of each cohort's offending as attributed to each offending classification group. The table shows close similarity between the 'single' and 'repeat' groups and some differences between these cohorts and serial perpetrators, with the latter committing proportionately more criminal

damage and 'other' domestic abuse crimes. The differences in these proportions that are unlikely to be due to chance are denoted by asterisks in the table. They are between single-time and serial perpetrators for violence with injury (*t (2265) = -2.372, p = .0089)*, criminal damage (*t (2961) = -6.019, p =0.00)* and rape (*t (3908) = -5.351, p = .00;* and between repeat perpetrators and serial perpetrators for violence with injury (*t (3794) = -1.949, p = .0257)* and rape (*t (4104) = -3.939, p = .00)* and so our interpretation of these data is that serial perpetrators have a greater tendency toward broader offending.

It is also worthwhile considering these data from the perspective of the contribution of each cohort to the level of each type of crime. Overall, serial perpetrators (which made up 10% of all perpetrators in the database) accounted for 21% of domestic crimes. In comparison, 26% of crimes were attributed to repeat offenders (who accounted for 15% of all offenders) and 53% of crimes were attributed to single-time offenders (who were 75% of all offenders). We would not expect an equal distribution, as single-time offenders by definition have at least one less crime than every repeat or serial offender. This pattern does not, however, hold for all crime classifications. Repeat offenders contributed almost twice as many rape crimes as serial offenders (28% to 15%), whereas 28% of domestic abuse criminal damage offences were attributed to serial offenders – 3% more than to repeat offenders.

16.4 Harm

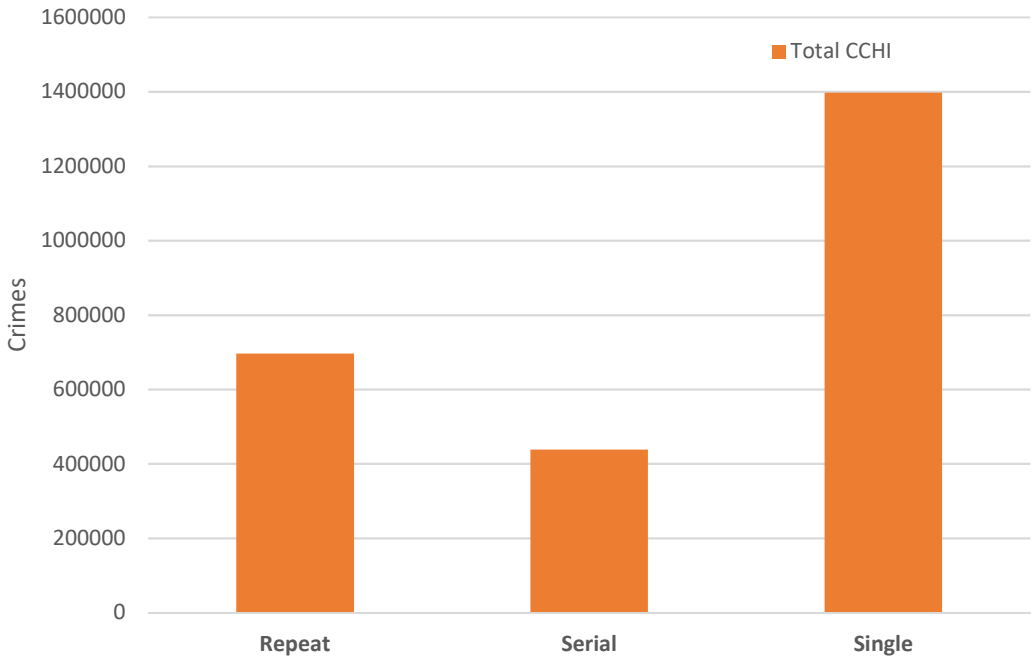

Figure 12. Total crime and crime harm by cohort

Figure 12, which shows the relative contributions to crime count and total crime harm, demonstrates that most harm (a total of 55%) is attributed to single-time perpetrators, but this perspective needs to be considered alongside the number of individual perpetrators in each cohort. Figure 13 shows the mean level of crime harm per offender in each cohort – a view which clearly indicates that serial offenders of domestic abuse accounted for more than twice as much harm per offender than single-time offenders (which makes sense because, by definition, there will be at least twice as many crimes attributed to each serial perpetrator compared to single-time perpetrator), but slightly less harm than repeat offenders.

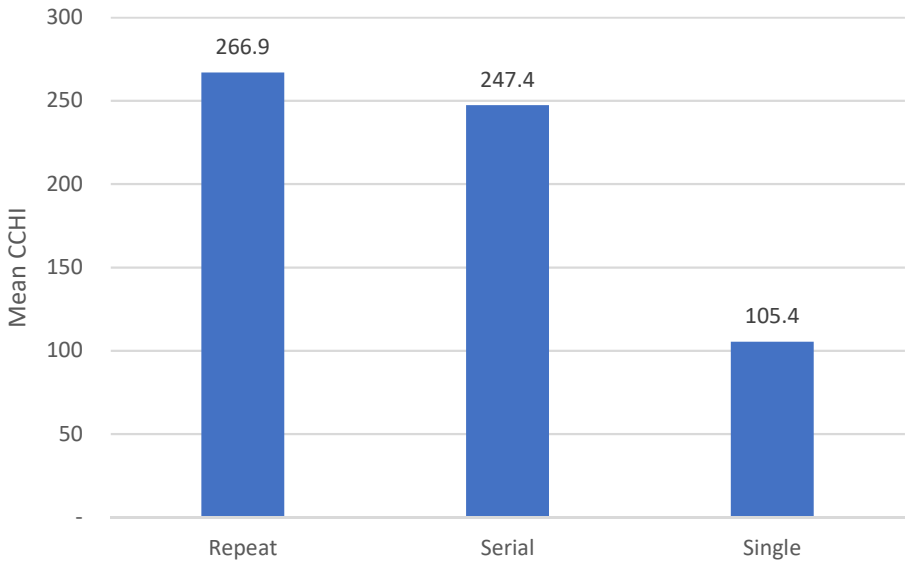

Figure 13. Mean CCHI per offender per cohort

In terms of the 'power few' concept, which is explored in more detail in Chapter 18, Figure 14 indicates that 80% of total harm in this dataset was attributable to just 6% of perpetrators (1,081 individuals).

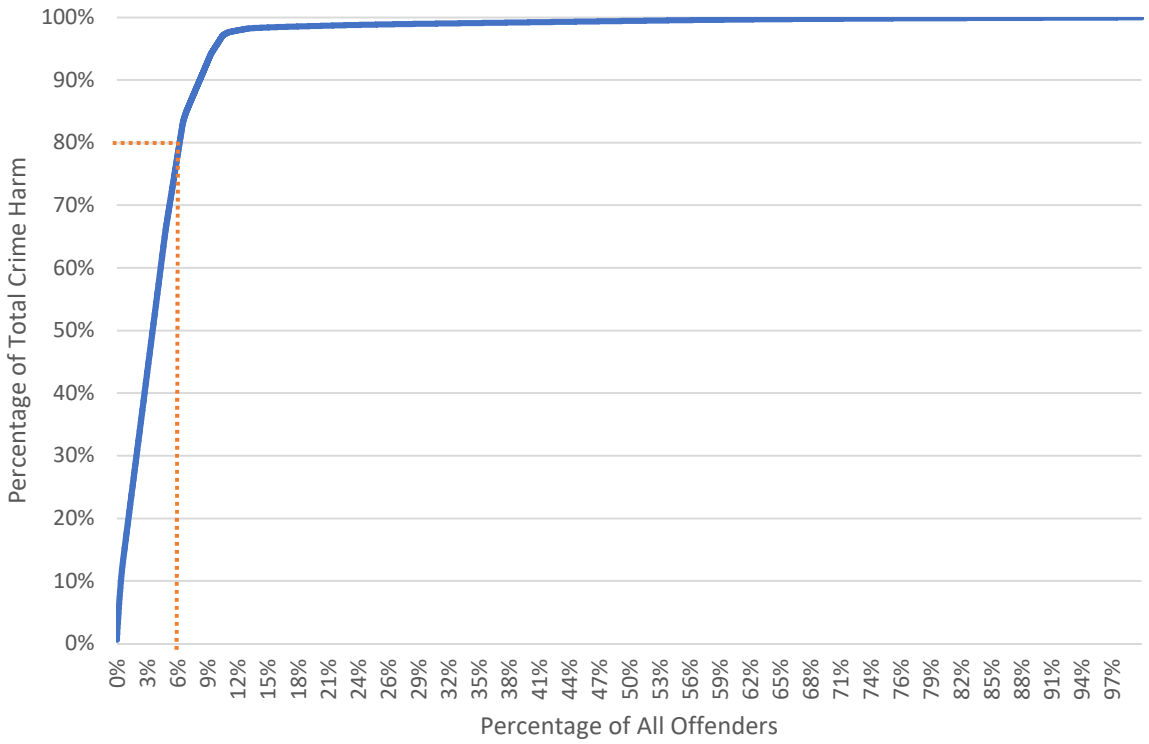

Figure 14. Power curve graph for cumulative proportion of crime harm by cumulative proportion of offenders

Among the 1,081 perpetrators in the 'power few', 17% were classified as 'serial', compared to 10% in the database overall. Repeat offenders, which made up 15% of all offenders, composed 27% of the 'power few', and the remaining 56% were single-time offenders. Overall, repeat or serial offenders were twice as likely as single-time offenders to form part of the 'power few' cohort. These figures are shown in full in Table 23.

Table 23. Power few contributions of different offender cohorts

Offender type	Contribution to power few	Proportion of total who were in the power few
Single	56%	5%
Repeat	27%	11%
Serial	17%	10%

16.5 Other Crimes

There were 147,512 non–domestic abuse crimes committed within the follow up period in the police force under examination, 25,302 (17%) of which were linked to one of the domestic

abuse offenders in our database, via either “suspect status” or “formal sanction”[18]. These crimes were distributed among 7,079 (40%) of the 17,641 domestic abuse offenders in the dataset.

[18] Suspect status is assigned when an individual is suspected by investigators of perpetrating an offence. A formal sanction is applied when the investigation concludes that a suspect did commit the offence. Sanctions take multiple forms including charges (which are referred to a court), cautions or fixed penalty notices.

Table 24 shows the breakdown of prevalence of non–domestic abuse offending among the three perpetrator cohorts.

Table 24. Prevalence of non–domestic abuse offending among cohorts

Crime class	Single	Repeat	Serial
Arson and criminal damage	6%***	12%***	20%
Burglary	2%***	5%***	8%
Drug offences	5%***	12%***	16%
Miscellaneous crimes against society	2%***	5%***	8%
Possession of a weapon	1%***	3%***	6%
Public order	8%***	17%***	25%
Robbery	1%***	2%***	4%
Sexual crimes	2%***	5%**	7%
Theft	6%***	12%***	19%
Vehicle crimes	2%***	3%***	5%
Violent crimes	21%***	40%***	51%
Statistical significance in difference compared to serial perpetrators ** p< .05, ** p<.01, *** p<.00*			

This presents a stark (and according to t-tests of proportions, statistically significant) difference between the cohorts. In every category of crime, a greater proportion of serial perpetrators were linked to non-domestic offending in the time window. It is also notable that repeat offenders were proportionately more often linked to non-domestic crimes than single-time domestic abuse offenders. In total, 70% of serial perpetrators (1,233 of the 1,770) were linked to non–domestic abuse crimes, compared to 57% of repeat offenders and 33% of single-time offenders, suggesting a greater tendency toward generalist offending among the serial cohort which we will return to in Chapter 12.

These trends also extend to the measurement of harm, as shown in

Table 25. Serial perpetrators were more likely to be linked to higher-harm non-domestic offences than non-serial offenders, to a statistically significant level, according to independent *t*-test comparisons between serial and repeat (t (2,703)= 2.33, p = .0203) and between serial and single (t (5,575) = 5.34, p = .0001) offenders.

Table 25. Mean CCHI of non–domestic abuse offending among cohorts

	Single	Repeat	Serial
CCHI of non–domestic abuse offences (*M*)	256.5***	317.8**	391.1
(*SD*)	757.2	781.2	858.5
Number of offenders linked to non–domestic abuse offences	4,348	1,476	1,229
Proportion of offenders linked to non-domestic abuse offences	33%	57%	70%

Statistical significance in difference compared to serial perpetrators
** p< .05, ** p<.01, *** p<.00*

This variance originates primarily in violent crimes, in keeping with the results shown in Table 22. Figure 15 shows that serial perpetrators accounted for higher average harm in higher-harm violence and offences related to the possession of weapons.

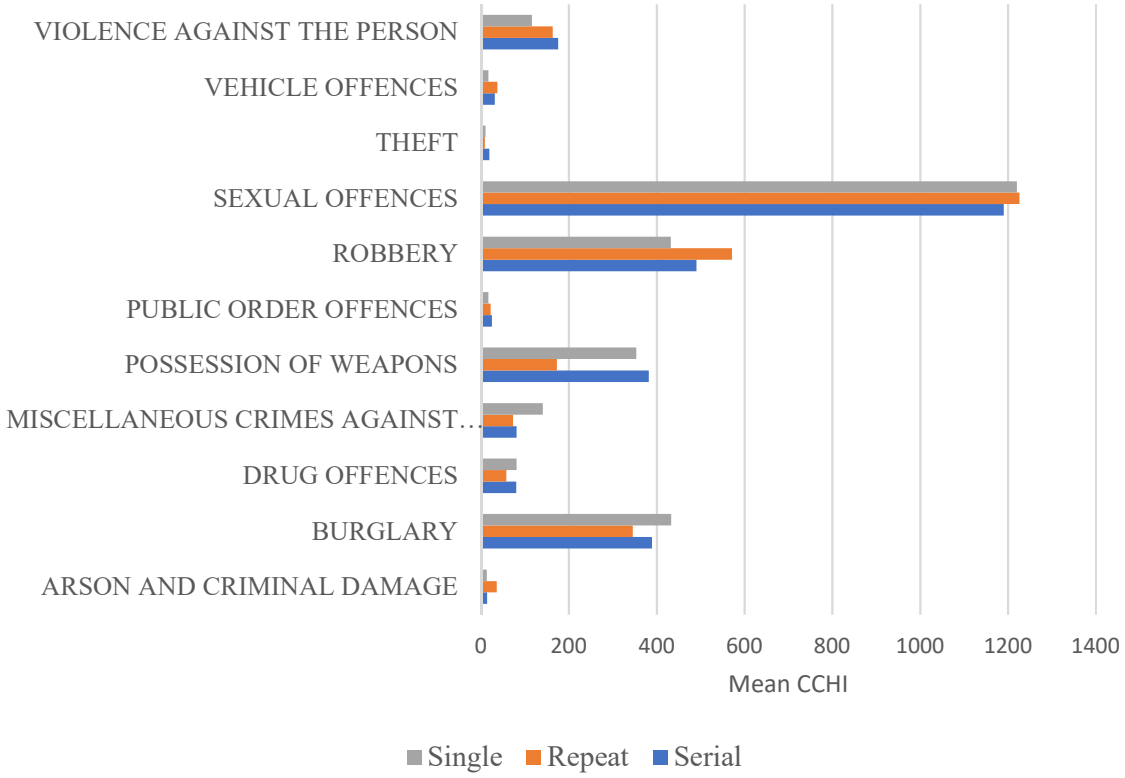

Figure 15. Average non–domestic abuse CCHI by crime type and cohort

16.5.1 Subclassifications of cohorts

As we have seen so far, serial perpetrators offended more frequently and more harmfully with respect to non–domestic abuse crime in the same follow up period as the dataset. This apparent tendency towards generalist non-domestic offending merits closer attention. Table 26 cross-references the domestic abuse cohorts we have analysed with three classifications of offending patterns, consistent with the typologies discussed earlier in the chapter (see Holtzworth-Munro and Stuart, 1994, etc.). These three categories are as follows: (1) ‘Family only’, which consists of only domestic offending; (2) ‘Violence offences only’, which consists of only domestic abuse and non-domestic violent crimes; and (3) ‘Generalist’, which includes any kind of offending.

Table 26. Mean domestic CCHI by cohort/offending type

Type	Family offences only	Violence offences only	Generalist
Single (M)	108.4***	80.6***	108.9***
(SD)	(342.7)	(353.3)	(396.6)
	n = 8,903	*n = 1,495*	*n = 2,863*
Repeat (M)	265.2**	217.2	289.3
(SD)	(611.9)	(536.3)	(418.4)
	n = 1,046	*n = 442*	*n = 1,046*
Serial (M)	193.0	172.7	300.7
(SD)	(361.8)	(427.3)	(212.0)
	n = 537	*n = 283*	*n = 950*
Statistical significance in difference compared to serial perpetrators in the same category (e.g. family only) ** p< .05, ** p<.01, *** p<.00*			

Table 26 displays the mean domestic CCHI value for these nine new sub-classifications. While it repeats some of the information shown earlier in this Chapter (e.g., that repeat and serial offenders are on average more harmful), the cross-section reveals that, among the serial cohort, the 'Generalist' serial perpetrators were the most harmful, by more than 50%, compared to other serial perpetrators, with two-way t-tests indicating statistically significant differences between single-time and serial offenders in each category, and between repeat offenders (more harmful) and serial offenders in family only violence. Otherwise there was no significant difference between repeat and serial domestic abusers. The link between harm and generalist offending is further extended by the fact that 'Generalist/Repeat' is the second most harmful sub-categorisation. We discuss these trends further in Chapter 20.

16.6 Summary

Datasets 1 and 2 showed similar prevalence levels for serial perpetrators – between 10% and 15%. There were few statistically significant differences in the demographic composition of this cohort and repeat perpetrators, but the analysis found that serial perpetrators tend to be marginally younger than single-time offenders.

The profile of domestic crimes committed by serial perpetrators was slightly different to other groups but the mean level of harm attributable to a serial offender was comparable to that of a repeat offender. However, there is a strong trend among serial perpetrators for more volume and more harm in other types of non-domestic crime, of every kind. Serial perpetrators have been shown to be more ‘generalist’ than the other two cohorts, and those serial perpetrators who demonstrated this offending pattern were the most harmful domestic offenders overall, alongside repeat generalist offenders.

17 Escalation Findings

17.1 Chapter roadmap

This chapter presents the results of analysis of Dataset 1 in respect of patterns of escalation. Recall that dataset 1 comprises multiple forces' data from multiple years – offering a greater chance of detecting temporal patterns than the shorter dataset 2, and the arrest-only dataset 3. Different subsections of this chapter deal with the respective differences between victims and offenders, with the same procedure: we plot the mean CCHI scores for each sequential crime and then use analysis of variance and Tukey's Honestly Significant Difference (HSD) tests to detect significant differences in harm. We find that for both groups, the analyses do not support the notion of general escalation of severity in police records. In fact, the most compelling statistical evidence is for a de-escalating effect after the first crime is reported.

17.2 Victims

Of the 170,391 victims in the dataset, 5.1% (8,704) were selected for analysis of escalating severity, as measured by each crimes CCHI score. Figure 16 shows the breakdown of total crime count frequency, illustrating the size of the sample overall. Due to the small sample sizes at higher frequencies, the analysis only evaluates the first 10 crimes of the offenders with 10 or more in total. Thus, the analysed sample size for each category is the total number of victims with five or more crimes (8,704) minus the total of those who did not reach the category total. For example – the sample size for the tenth crime is 8,704 minus the totals of the categories 5, 6, 7, 8 and 9 (because the victims with these totals did not have a tenth crime). Therefore, for the tenth crime n = 8,704 – 6,966 = 1,738.

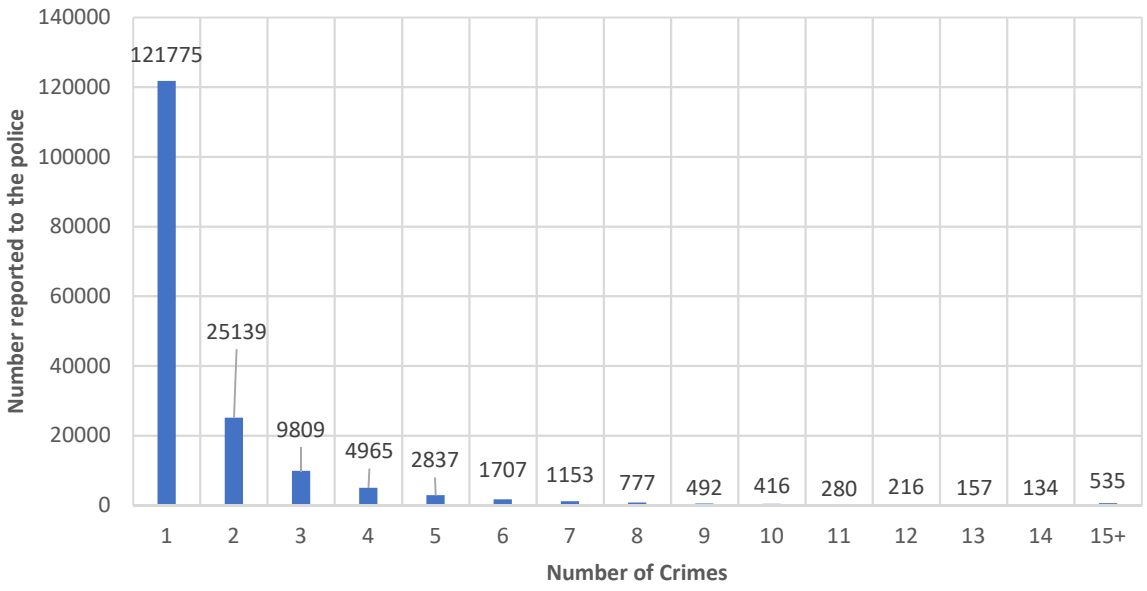

Figure 16. Sample sizes for number of total crimes for victims

Figure 17 shows that for these victims there was a general downward trajectory of the average CCHI score. A one-way ANOVA test determined the presence of statistically significant predictor for somewhere in the sequence ($F(9, 60{,}188) = 1.88$, $p \leq .001$), meaning that at least one point, the result was not due to chance alone. Given that two of the sequential crimes (5 to 6 and 8 to 9) gave an immediate escalation of mean CCHI, and all crimes after the fifth were of higher mean CCHI than the fifth, the ANOVA result alone does not rule out the possibility that the statistical significance relates to a pattern of escalation (although mere visual inspection of the graph suggests that it is more likely to be associated to the difference between the first and subsequent crimes).

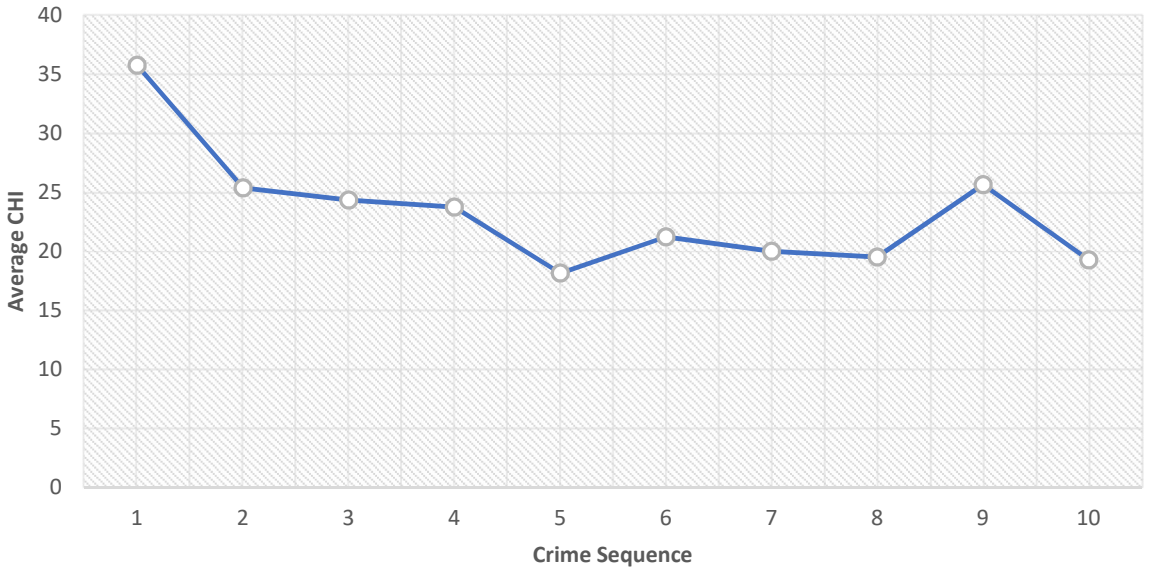

Figure 17. Average CCHI score over first 10 incidents for victims with 5+ crimes

A Tukey's HSD post-hoc test was applied to each result to determine where the significant difference(s) lay (see Table 27). These results indicated that the harm in the first incident was in fact statistically significantly higher than all others in the sequence, except the ninth. This contradicts the notion of escalation of severity after the first crime report.

Table 27. Tukey's HSD results for CCHI means attributed to victims with a minimum of five domestic abuse events

INCIDENT	1	2	3	4	5	6	7	8	9
1	–								
2	10.81*	–							
3	11.38*	0.57	–						
4	10.54*	0.27	0.84	–					
5	12.89*	2.08	1.51	2.35	–				
6	9.67*	1.15	1.72	0.87	3.23	–			
7	10.85*	0.03	0.54	0.31	2.05	1.18	–		
8	11.95*	1.14	0.57	1.41	0.94	2.28	1.10	–	
9	5.23	5.58	6.15	5.31	7.66	4.43	5.61	6.72	–
10	11.63*	0.82	0.24	1.09	1.26	1.96	0.78	0.32	6.39

*Note: Critical range = 7.68; * $p < .05$; ** $p < .01$; $p < .001$*

17.3 Offenders

As covered in Chapter 15 and mirroring the victim profile, the majority of offenders were not repeatedly linked to domestic abuse crimes in the multiyear period covered by the dataset. But of the 155,590 offenders in the data, 6% (9,337) - a greater proportion of offenders than victims - were linked to five or more crimes and thus analysed for patterns of escalation. Figure 18 shows the full breakdown of total crime counts among offenders in the database. Despite a greater overall tendency to repeat abuse among offenders (see *Repeat Abuse Findings*), there were fewer offenders than victims in each total crime category after five crimes.

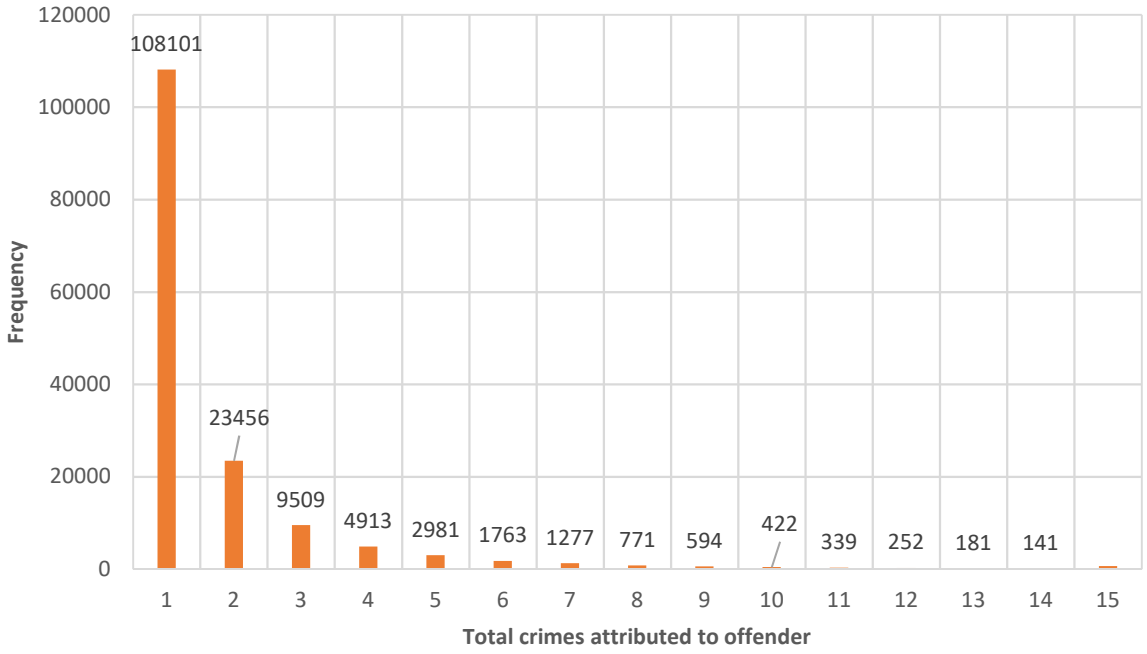

Figure 18. Sample sizes for number of total incidents for offenders

Figure 19 shows the pattern of mean CCHI across those first 10 crimes. The pattern mirrors the same analysis for victims in that the first recorded crime has the highest mean harm. However, after the second crime there is an observable pattern of increase up to the sixth crime, followed by another chronological pattern of rising severity between the seventh and ninth. A one-way ANOVA test for offenders determined that there was a statistically significant difference in at least one pairing in these data ($F(9, 55{,}632) = 1.88$, $p \leq .001$).

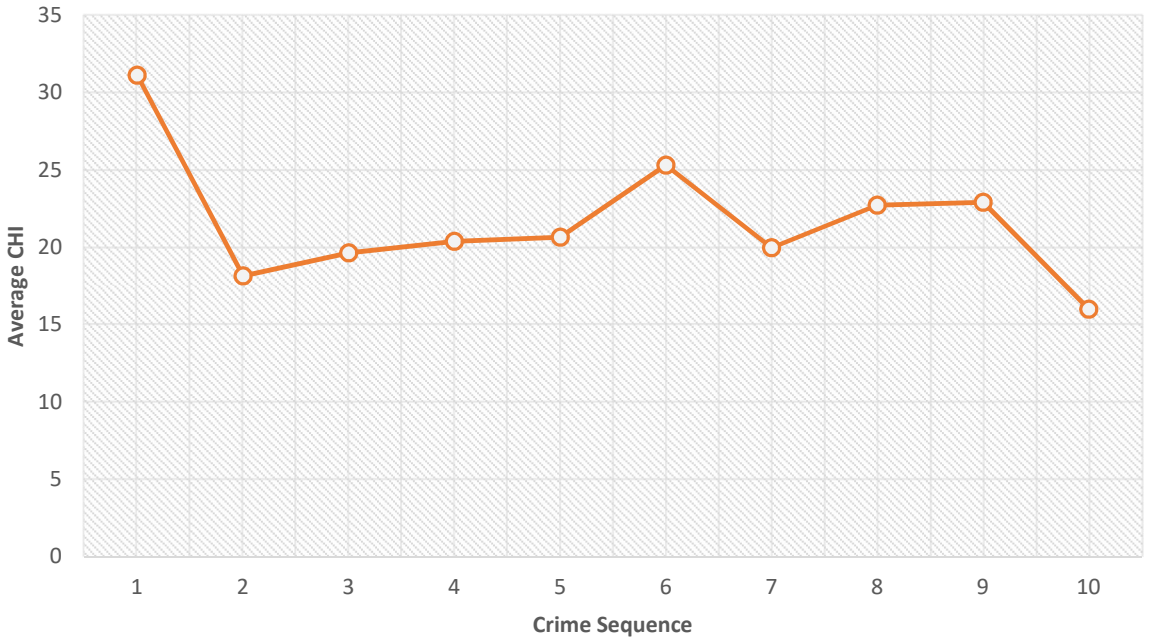

Figure 19. Average CCHI score over first 10 incidents for offenders with 5+ crimes

To detect where the statistical significance lies within these crimes, Tukey's HSD was undertaken, just as for victims (see Table 29 for the full results). Just as for victims, this procedure found that the first recorded crime was significantly different, although for offenders this was without exception. There was also a significant difference between the sixth and the 10th.

Table 28. Tukey's HSD results for CCHI means attributed to offenders with a minimum of 5 domestic abuse events

INCIDENT	1	2	3	4	5	6	7	8	9
1	–								
2	36.73*	–							
3	38.03*	1.30	–						
4	38.20*	1.47	0.17	–					
5	38.36*	1.63	0.34	0.16	–				
6	33.74*	2.99	4.28	4.46	4.62	–			
7	39.24*	2.50	1.21	1.03	0.87	5.49	–		
8	36.58*	0.15	1.45	1.62	1.79	2.83	2.66	–	
9	36.37*	0.36	1.65	1.83	1.99	2.63	2.86	0.20	–
10	43.33*	6.60	5.31	5.13	4.97	9.59*	4.10	6.75	6.96

*Critical range = 8.35; * $p < .05$; ** $p < .01$; $p < .001$*

17.4 Summary

For both victims and offenders, these analyses suggest that theories of escalating harm or severity are not borne out by police records. In fact, de-escalation of harm is largely present after the first recorded crime and there is no difference in harm after that. The implications of these findings are discussed further in Chapter 20.

18 Concentrations of Harm Findings

18.1 Chapter roadmap

This chapter presents the findings of analysis of Dataset 1 in relation to research questions 13-16. It is divided into two subsections – one for the measurement of the 'power few' for both victims and offenders, and one for the analysis of 'never called before' cases – members of the 'power few' with only one crime in the dataset. Using CCHI, the results show disproportionate levels of concentration of harm among both victims and offenders and illustrate the extent to which police have no prior domestic abuse records for the 'power few' before serious harm occurs.

18.2 Power Few

Table 29 illustrates the concentrations of CCHI scores among a small proportion of both victims and offenders. Across the four forces in the dataset, 80% of crime harm was attributed to fewer than 5% of the individuals involved. The high number of individuals associated with comparatively low CCHI scores was reflective of the high proportion of 0–30 CCHI events within the dataset – 96% of crimes had CCHI scores lower than 30 days. The mean CCHI for all victims was 52.1 days. By contrast the mean in the 'power few' was 1,553 days, and the minimum 549.2 days. For offenders the overall mean of total CCHI was 56.3 days, for the 'power few' offenders it was 1,611 days and the minimum was 551.25 days. These minimum levels translate to a particular set of crimes which distinguish the threshold for entry into the proportionally small cohorts which contribute most harm. Of the crime classifications which are weighted at 548 days, the most commonly occurring in the dataset were grievous bodily harm without intent, false imprisonment and kidnap. For further context, in the top 25% of harm, victims and offenders had to be linked to at least one crime of rape, attempted homicide, or arson with intent to endanger life. A homicide would automatically be included too.

Table 29. Number of domestic abuse crimes in dataset attributable to highest-harm offenders and victims

Cumulative percentage of total crime harm	Number of victims	Cumulative percentage of total victims	Number of offenders	Cumulative percentage of total offenders
5%	80	0.1%	79	0.1%
10%	250	0.1%	208	0.1%
25%	928	0.5%	1,276	0.8%
50%	2,145	1.3%	2,066	1.3%
80%	4,605	2.7%	4,349	2.8%
100%	170,391	100%	155,490	100%

Table 30 shows a comparison between the victims and offenders within the respective 'power few' (2.7% victims and 2.8% offenders contributing 80% harm) and non 'power few' groups (97.3% of victims and 97.2% of offenders contributing 20% of harm). For both victim and offenders, the 'power few' were older and less frequently male than those outside the 'power few' with statistically significant results from t-tests of means and proportions.

Table 30. Demographic comparisons between 'power few' and non-'power few' victims and offenders

Characteristic	Victims		Offenders	
	Power Few	Non-Power Few	Power Few	Non-Power Few
Age (M)	27.9***	26.6	26.6***	23.4
Age (SD)	16.1	18.4	16.3	18.1
% Male	14.4***	15.6	61.0***	70.3
Comparison of Power Few to non-Power Few **** p < .05; ** p < .01; p < .001***				

18.3 Never called before (or again)

Of the 4,605 victims who were attributed to events accounting for 80% of all the domestic abuse harm, 1,904 (41%) featured just once in the dataset, indicating no record of domestic abuse before or after the serious crime. Although the 59% with multiple calls represents a

higher level of repeat victimisation among the 'power few' victims than the general victim population (indeed, in the 'power few' repeat victims are in the majority whereas, as described in Chapter 15, in the data overall repeat cases were 25%) 41% still represents a significant proportion of serious cases in which police had no prior domestic opportunity to intervene. To compound this further, among the repeat cases in the 'power few' cohort, there was limited opportunity to forecast and prevent harm – 78% of the cohort (including the single-time victims) had fewer than five crimes in the dataset.

Of the 4,349 offenders linked to 80% of harm, 1,710 (39%) appeared in just one record. The majority of high harm offenders – 73% - had fewer than five crimes. The actual window of forecasting is probably even more limited, because it is improbable that the 'serious' offence – the one which 'qualified' the offender for 'power few' status, was the last one in the every sequence. Indeed, of the 7,041 crimes in the dataset which had a CCHI score of more than 548 (the nominal 'power few' threshold outlined in the previous section), 67% and 69% were the earliest recorded crime for victims and offenders respectively. Just 8% of serious crimes occurred later than the fourth sequential crime both victims and offenders.

18.4 Summary

These findings suggest that the majority of serious harm domestic abuse cases is limited to an extremely small proportion of the overall population of victims and offenders. Any crime classified as kidnap, false imprisonment, grievous bodily harm (with or without intent), rape, arson endangering life or homicide (including attempted homicide) constitutes 'power few' status. Among the victims and offenders linked to such crimes, however, around 40% have no prior record of domestic abuse, and around three quarters have, at best, an extremely limited opportunity to predict their serious harm status, emphasising the need to seek methods of predicting high harm cases without relying just on prior records.

19 Forecasting Findings

19.1 Chapter roadmap

This chapter presents the results of analysis in relation to research questions 16-20, concerning the application of a random forest forecasting model to a dataset comprised of arrest records for any type of crime. The chapter begins with the answers to questions 16 and 17, regarding the baseline measures for the model – that domestic abuse is a relatively rare occurrence as a proportion of all arrests, that serious domestic abuse is especially rare and that based on a 24-month time horizon, the forecasting model could, at best, identify almost half of all 'serious' domestic abuse arrests but with high levels of 'true negatives'.

The chapter then presents the results of the forecasting model, which is deliberately skewed towards cautious errors, with an abundance of 'false positive' forecasts (where serious domestic abuse is predicted but none occurs) but a very high accuracy at predicting serious abuse overall and an almost perfect record when no abuse is forecast (true negatives). The model is universally more accurate than the baseline and would predict more than three quarters of the serious domestic arrests that have some form of prior arrest record, which equates to more than a third of all serious domestic arrests that the police could have an opportunity to intervene with. The final section presents information about the predictor variables with the greatest influence on accuracy.

19.2 What proportion of all arrestees go on to commit domestic abuse?

As outlined in Chapter 14, one of the first steps to take in any forecasting procedure is to determine the 'baseline' level of the outcome that is the subject of the forecast. In the simplest terms, this can be presented as an answer to the rudimentary question: 'If all forecasts were assigned to just one outcome classification, what proportion would be correct?' In the model under scrutiny here, there are three possible outcomes to assess, the most frequent of which is 'no domestic abuse' within the two-years-from-arrest time horizon. As Table 31 shows, 79.5% of arrestees were not arrested for any form of domestic abuse in the follow-up period. In practical terms, were the model to forecast that no domestic abuse would occur in every case, it would be correct in around eight out of every ten cases. It would also universally fail in its objective, because despite having a high overall degree of accuracy, it would not predict any future serious crimes. This serves to demonstrate just why overall

accuracy is not the best determinant of model performance, but rather a full breakdown of true and false negatives and positives.

Conversely, if every prediction were for serious abuse to occur, the model would be *incorrect* 99.1% of the time and 80% of the time if every forecast were of less-serious domestic abuse.

Table 31. Baseline levels for domestic abuse outcomes

Category	Percent
Proportion of arrestees with no domestic abuse arrest within 24 months of arrest	79.5%
Proportion of arrestees with a less-serious domestic abuse arrest within 24 months	19.6%
Proportion of arrestees with a serious domestic abuse arrest within 24 months	0.9%

19.3 What proportion of domestic abuse arrestees have prior domestic records?

As we have seen in Chapter 18, police data indicate that perpetrators of serious domestic abuse tend to have no or little prior domestic abuse record in a substantial proportion of cases - indeed, this is one of the main reasons for exploring the entire population of arrestees for forecasting potential. Nevertheless, examining the extent to which domestic abuse offenders have prior domestic arrests in the datasets under examination here is an important contextual point for establishing external validity as well as for helping to interpret the resulting model.

Table 32 presents a range of descriptive statistics about the sample cross-referenced by the actual outcome and uses t-tests (including t-tests for proportions) to determine the significance of differences.

Table 32. Profile of cases in training dataset

Indicator	No arrest	Less serious arrest	Serious arrest
Number of arrestees by outcome for two years after the index arrest	58,301	14,398	681
Proportion with arrest record prior to the index arrest	80.9%***	87.9%***	94.3%
Proportion with domestic arrest record prior to the index arrest	21.1%***	44.1%**	48.6%
Proportion with serious crime arrest record prior to the index arrest	24.0%***	28.2%***	44.6%
Proportion with violent crime arrest record prior to the index arrest	59.1%***	87.9%***	84.0%
Number of priors (*M*)	16.9***	16.8***	25
(*SD*)	(30.9)	(26.8)	(31.5)
Age at first arrest (*M*)	23.9***	22.2	21.8
(*SD*)	(11.9)	(11.5)	(11.7)
Proportion male	83.5%***	90.8%	87.1%
**=p< .05, ** = p< .01, *** = p< .001 – significance compared to future serious DA arrest (DV2)*			

Logically, for any statistical model to accurately forecast serious domestic cases as distinct from less-serious ones, there must be patterns of differences in the predictor variables. These predictors demonstrate enough statistically significant differences between the outcome types (particularly no arrest and arrest for a serious domestic crime) to suggest that modelling may be possible. The following differences should be noted:

- Subsequent serious arrestees had a prior arrest record *more often* than arrestees who later committed less-serious offences (Odds ratio (OR) = 2.3 (95% confidence interval (CI) = 1.6-3.1) or no abuse (OR = 3.8, CI = 2.8-5.4).

- Subsequent domestic arrestees (both serious and less-serious) more commonly had prior arrests for domestic abuse than those who committed no further abuse (Serious OR = 3.5, CI = 3.0 – 4.1; Less-serious OR = 2.9, CI = 2.8 – 3.1).

- Less-serious domestic arrestees more commonly had a prior record for a serious crime than those going on to commit no domestic abuse (OR = 1.2, CI = 1.2 – 1.3), and subsequent arrestees for serious domestic abuse had such a prior record even more commonly (OR = 2.5, CI = 2.2 – 3.0).

- In equal measure, subsequent domestic arrestees (both serious and less-serious) more frequently had priors for violent crime (Serious OR = 3.6, CI = 3.0 – 4.5); Less-serious OR = 5, CI = 4.8 – 5.3). Strikingly, almost nine in every 10 arrestees who went on to be arrested for domestic abuse within two years had some kind of prior arrest record for a violent crime.

In respect of serious domestic abuse in particular, these results are of notable interest. In Chapter 18, it was established that around half of serious domestic abuse crimes involved perpetrators with no prior domestic record (in a time limited period). These data support this but add a striking new aspect to the analysis – that these offenders probably had some kind of prior record for another form of crime, and it was probably a violent crime. This does, however, raise a fundamental question: What proportion of serious domestic arrestees have any form of prior arrest record *within the preceding 24 months*? If the earlier analysis regarding prior domestic records translated to no prior records at all in around half the cases, and of the approximate remaining half, only a small proportion had that prior arrest within two years, then the potential impact for the forecasting model would be very limited because the police would not be able to use arrest data as a potential source for identifying these offenders. As it happens, this is not the case.

Table 33 shows that almost half of all serious domestic abuse arrests would be exposed to the forecasting model if configured on a 24-month follow-up period because they had been arrested for any form of crime within the two years prior (i.e. if the offender arrested for a serious domestic crime had no prior arrest record in the preceding two years –

the model would not be able to forecast them because there would be nothing to trigger it). This rose to 59% of less-serious domestic abuse arrests.

However, because we had access to more than ten years of arrest data, we were able to assess the impact of extending the window to more than two years. If the follow-up period is extended, these proportions increase to the point where over three quarters of cases would be exposed to the model in a 20-year follow-up. This is clearly impracticable; police forces do not have 20 years of reliable domestic abuse records on which to train such models, but this will not always be the case. These implications are discussed further in Chapter 12.

Table 33. Proportion of domestic abuse arrestees with prior arrest records (for any type of crime)

Follow-up period (months)	Less-serious	Serious
24	59.0%	48.7%
36	64.9%	56.7%
48	68.9%	60.1%
60	71.6%	62.5%
120	78.6%	71.3%
240	82.0%	75.7%

19.4 Can antecedent inputs predict future domestic abuse cases to a high degree of accuracy?

.

Table **34** presents a summary of the output of the forecasting model based on the training dataset. The summary is presented in the form of a 'confusion matrix', which is the standard format for communicating the output of a forecasting model (see Berk, 2012).

Table 34. **Summary table for forecasting model accuracy**

	Forecast				
	No DA	Less serious DA	Serious DA	**Total**	*Outcome Class Accuracy*

		a			c			
	No DA	48,229	b	7,075	2,997	**58,301**		*83%*
Actu	Less serious	d			f			
al	DA	611	e	12,254	1,533	**14,398**		*85%*
		g			i			
	Serious DA	16	h	142	523	**681**		*77%*
	Total	**48, 856**		**19,471**	**5,053**	**73,380**		
Forecast Accuracy		*99%*		*63%*	*10%*			

Table 34 is a cross-tab, with each of the main boxes (labelled 'a' to 'i') cross referencing forecast results for every record in the data against the actual outcome. Each record in the data has been processed using the forecasting model so the totals are equivalent to the overall sample size (n = 73,380). The forecasts are organised into the three categories of outcome: no DA arrest, less serious DA arrest and serious DA arrest. Where a forecast of No DA arrest was made, and the actual outcome was no DA arrest, the results are stored in box 'a'. This cell is shaded green because the forecast was correct. The same shading is also applied to where less serious DA arrest forecasts were actually less serious DA arrests ('e') and arrests for serious DA forecasts actually arrests for serious DA ('i'). Added together these three cells represent the overall accuracy of the model (48,229 + 12,254 + 523). Thus, of the 73,380 forecasts made 83.1% were correct – surpassing the baseline accuracy threshold of 80% (see *Baseline considerations*).

However, the interpretation of accuracy is more nuanced than this. We are primarily concerned with the accurate prediction of serious domestic abuse, and the model was 'tuned' to be more accurate in respect of this outcomes than the others. The result is a high rate of 'false positive' or 'cautious errors'. Consider the Serious DA forecast column: overall 5,053 forecasts of serious domestic abuse were made, yet just 10% turned out to be correct. Compared to the percentage of forecasts for no domestic abuse and less-serious domestic abuse (99% and 63% respectively) this seems like a poor result, especially given the tuning which made the model more accurate at predicting serious harm. However, we must also consider that serious domestic abuse is very rare – there were only 681 instances of it in a dataset of more than 73,000 records. The model correctly identified 77% of these. So, while 90% of serious abuse forecasts may not actually result in that outcome, the 10% that do,

account for more than three quarters of all the serious domestic abuse arrests in the subsequent two years where the offender had a prior arrest record within two years. Given that, in the time horizon analysed (two years), 48.7% of serious domestic arrests had some kind of prior arrest, this means that this model would have correctly identified 37% of all serious domestic abuse. In Chapter 20 we consider whether this is good enough.

Model accuracy is not completely summarised in the rate of successful prediction of serious abuse, however. A more complete breakdown of model performance is given in Table 35. One of the key concerns is the number of 'dangerous' forecasts – those where no abuse or less-serious abuse is forecast, and serious abuse subsequently occurs. Given that 77% of serious abuse was successfully identified by the model, naturally 23% was not. But how confident could users be that when the model makes a forecast of no abuse or less-serious abuse, that it will not result in such a 'dangerous' error. The answer seems to be 'even more confident'. In 99% of the cases that were forecast to have no domestic abuse in two years, the forecast was correct. When less-serious domestic abuse was predicted, almost two thirds of forecasts were correct and of those that were not, fewer than 1% were because a serious crime occurred. Table 35 shows the full summary of the model's performance.

Table 35. Model performance

Total proportion of forecasts that were accurate within two years	83.1%
Of those forecast to be arrested for a 'serious' domestic crime, percentage who actually were arrested for a 'serious' domestic crime	10.4%
Of those forecast to be arrested for a 'less-serious' domestic crime, percentage who actually were arrested for a 'less serious' domestic crime	62.9%
Of those forecast not to be arrested for any domestic offence, percentage who actually were not arrested for DA of any kind	98.7%
Of serious domestic offences, percentage correctly forecast	76.8%
Of less-serious domestic offences, percentage correctly forecast	85.1%
Of no domestic offences, percentage correctly forecast	82.7%
Cautious error to dangerous error[19] ratio:	15:1

[19] Cautious errors refer to those forecasts for serious and less serious domestic abuse which were actually no abuse (for either) or less serious abuse (for serious forecasts). Conversely, dangerous forecasts refer to forecasts of no abuse which actually were less serious or serious abuse and less serious forecasts which were actually serious.

Very cautious error[20] to very dangerous error ratio:	187:1
Proportion of no arrest forecasts which actually were an arrest for a serious offence	0.03%
Proportion of less serious DA arrest forecasts which actually were an arrest for a serious offence	0.07%

Based on these tables, we draw several conclusions about how the model performed. More than three quarters of forecasts of any kind were correct, and of those that were not, almost all the errors were 'cautious' – where the forecast was for domestic abuse to occur and it did not. In fact, the model comes extremely close to having no form of 'dangerous' error at all, which is both striking and unusual. Where the forecast was for no arrest for domestic abuse, it was almost always correct. When the model forecast an arrest for a less-serious abuse, it was correct more than half of the time, and correctly identified just over 80% of all less-serious domestic abuse.

19.5 Which predictors have the greatest impact on accuracy?

As described in *Predictor* in Chapter 6, we draw our conclusions about the influence of individual predictor variables from the R functions for variable importance plots and partial response plots. Recall that these show (1) the relative decrease in overall model accuracy were each variable removed in turn, and (2) how each variables values are related to the outcome classifications.

With 35 predictor variables in our model, we have a plethora of information available to us but for ease the results are summarised in Figure 20 and a selection of plots are subsequently included in *Appendix A: Technical Information Relating to Random Forest Modelling*. Figure 20 shows in descending order the relative importance to model accuracy based on mean accuracy scores. Accuracy scores are calculated for each tree in the random forest on a scale of 0 (least improvement) to 100 (most improvement).

[20] Very cautious errors refer to forecasts of serious abuse which were actually no abuse. Very dangerous errors refer to forecasts of no abuse when the actual result was serious abuse.

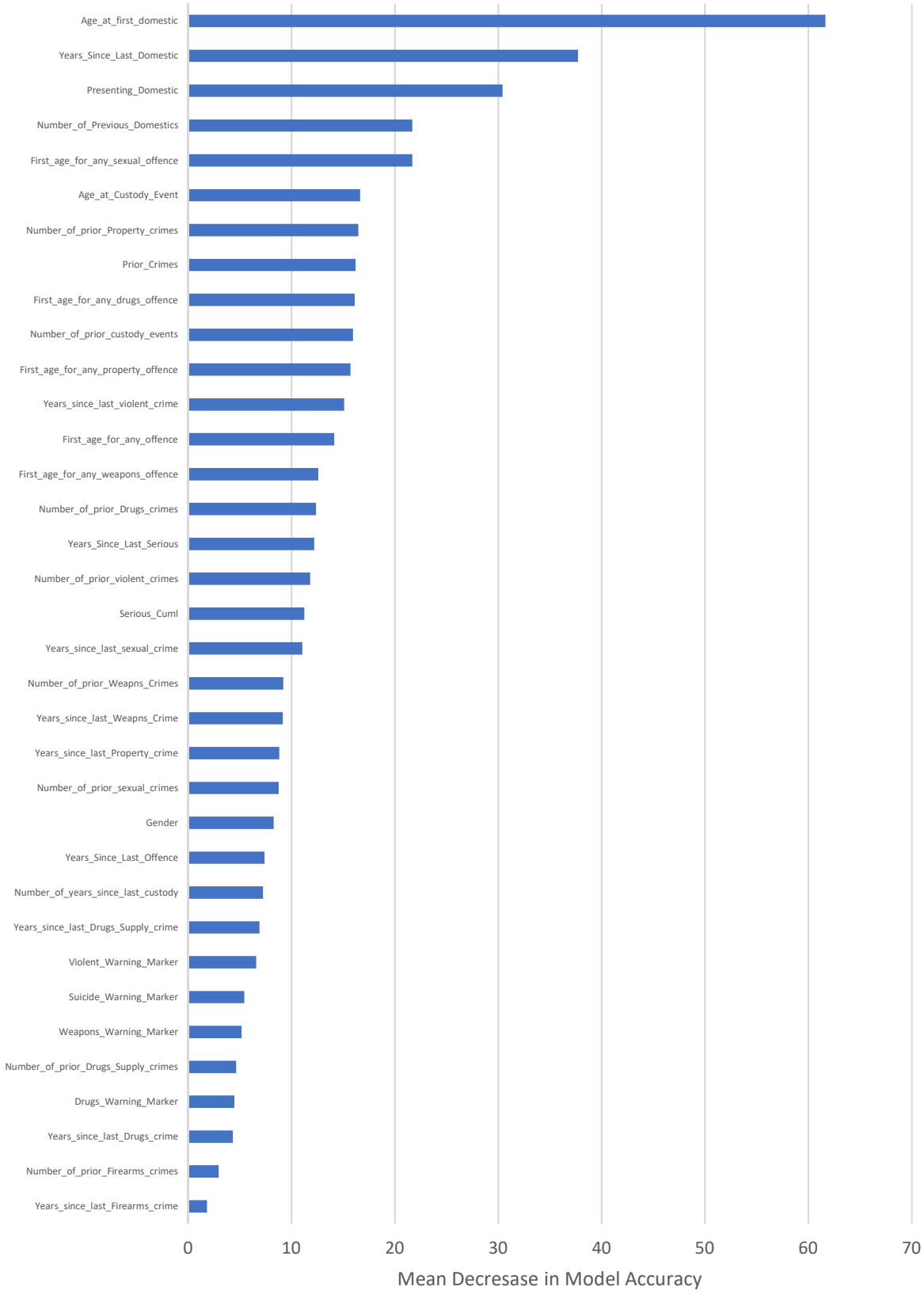

Figure 20. Variable importance plot for forecasting model accuracy

Figure 20 tells a clear story about the predictor variables relating to domestic abuse. 'Age at first domestic arrest' was consistently the most influential predictor variable in terms of model accuracy – 1.6 times more so than the next highest influencer. Predictors relating to whether the presenting arrest (on which the forecast was based) related to a domestic crime, and the number of years since the arrestee was last arrested for a domestic crime were the only other indicators to pass an average importance of 30. The total number of prior domestic crimes and the age at which an offender was first arrested for a sexual crime were the only other variables to surpass a mean of 20. This does not mean other variables did not contribute to overall accuracy – as Figure 20 shows, every predictor contributed something[21].

Figure 21 shows the respective levels of 'node purity' for each variable. As previously outlined, this refers to the extent to which the splits at each node) favour a particular outcome (no arrest, arrest for less serious domestic abuse or arrest for serious domestic abuse). For example, if a predictor variable was exclusively associated to no arrest it would have a score of 100 for no arrest and 0 for arrest for less serious or serious abuse. Figure 21 shows the mean purity (also known a 'Gini index') for all of the decision trees in which each variable was selected (recall that there were 501 trees in total).

[21] The thresholds selected in this section were drawn as natural breaks in the data.

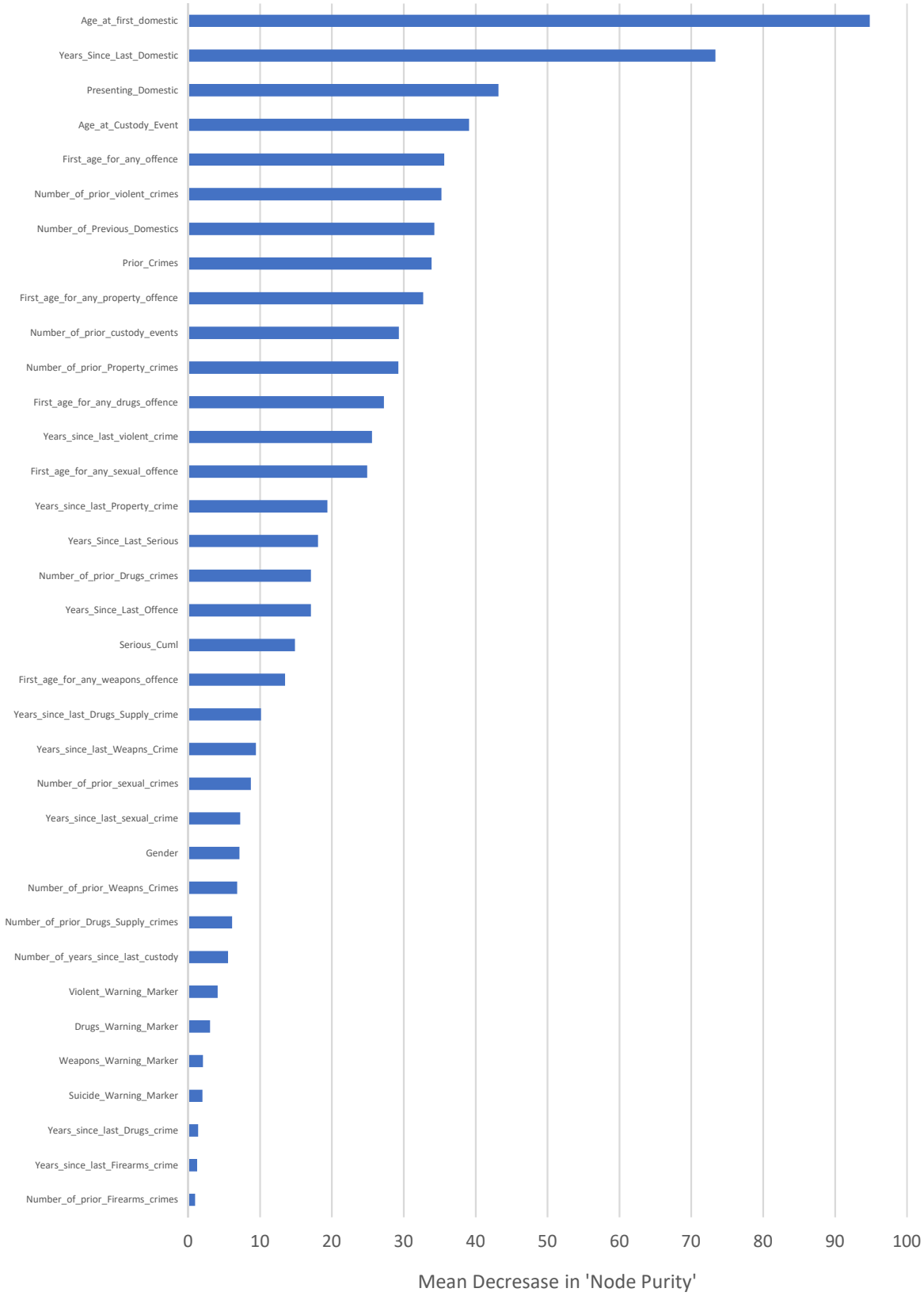

Figure 21. Variable importance plot for forecasting model node purity

The domestic predictor variables (prior number of domestic arrests and presenting arrest was domestic) were highly associated to particular outcomes but it is also notable that some variables rank more highly for 'node purity' than they did for overall accuracy contribution – i.e. they variable is correlated with a particular outcome, but not highly influential in the overall scheme of things. Figure 21 does not explain which outcome the variables are most associated with however – for this partial response plots are needed. Each variable can be cross referenced against all three outcome classifications (generating 105 plots overall). Plots for the top four variables for overall importance are included in '*Appendix A: Technical Information Relating to Random Forest Modelling*'. The results are summarised as follows:

19.5.1 Age first arrested for domestic abuse

This predictor variable is a proxy for having any prior record for domestic abuse arrest, because if the arrestee has no previous arrest for domestic abuse, they will have no corresponding age. Such cases were coded into the data as a value of –1 (random forests cannot process missing data, so any such data must be coded in some form[22]) and given the patterns shown in Table 32 it is perhaps not surprising to see some form of influence for this indicator. As Chapter 15 established, once linked once for domestic abuse, an offender is increasingly likely to be arrested again.

A '-1' value for age at first domestic arrest (i.e. there was no prior arrest) had the highest probability for forecasting no arrest and the lowest for less serious or serious domestic arrests. There were different patterns associated with first age and the forecasting of different types of abuse. Less-serious abuse within 24 months was most likely for arrestees who first presented for domestic abuse in their 20s. After this time, the likelihood reduced until arrestees are in their 50s. However, for serious abuse, the risk was greatest for first-time arrestees in their 40s.

19.5.2 Years since last arrest for domestic abuse

Like the previous predictor, the number of years since the last domestic abuse arrest is also a proxy measure for prior domestic arrests. Offenders with no prior domestic arrests (and therefore no years since their last) were coded as –1. As with the age of first domestic arrest then, we would expect to observe a higher probability of no abuse associated with this value.

[22] As our model was a classification model, coding this way did not affect calculations, as it would in a regression model.

Among these arrestees, it would be logical to expect that fewer years since the last domestic arrest were more likely to be associated with a higher probability of abuse, but it is perhaps difficult to predict, before knowing the results, whether there would be differences between less-serious and serious abuse.

As it turned out, just as with 'age at first domestic arrest', the absence of any data had the highest probability associated with no domestic arrest. Among those arrestees who did have a valid entry, the pattern was, as hypothesised: the greater the number of years since the last domestic arrest, the lower the probability of any form of future domestic abuse arrest. The probability of future abuse was at its highest at one year, but notably higher among less-serious cases than serious ones. The relationship between the predictor variable and both kinds of abuse outcome flatlined at around 10 years, which is when domestic abuse arrest record-keeping was of much lower reliability.

19.5.3 Presenting offence was domestic abuse

Cases in which the presenting arrest (the catalyst for the forecast) was tagged as a domestic crime were coded as a binary 0 (for no) or 1 (for yes). Cases which were not domestic had the greatest predictive association with no future abuse, as we might expect, and the inverse was true of presenting cases which were domestic. The association of a presenting domestic crime with less serious domestic abuse was however, more than twice as high as for serious abuse.

19.5.4 Number of prior domestic arrests

The pattern of previous domestic arrest history predicting future behaviour is replicated in the number of prior domestic arrests. Unsurprisingly given the proxy variables presented above, no prior arrests were most closely associated to a no future arrest forecast. Among less serious arrest forecasts there was almost no difference between the number of priors, with evidence of a slight decline in probability of less serious arrest being forecast the higher the number of priors. Forecasts of arrests for serious abuse were less associated with prior domestic arrests than less serious forecasts.

19.5.5 Age at first arrest for a sexual offence

Most arrestees did not have a prior sexual arrest and so were classified as '–1' values. The partial response analysis partially supports this, indicating a higher probability of a no arrest forecast among those with no prior sexual arrest record, and then a declining likelihood of no arrest until around age 40. This is complicated the pattern for less serious arrest forecasts, which shows a consistent downward trend in the likelihood of a less serious arrest the older

the arrestee was when first arrested for a sexual crime. This is almost entirely reversed for future serious arrests, which shows escalating likelihood of a serious arrest the older an offender was at the time of their first arrest.

19.6 Summary

The results presented in this chapter suggest that a domestic abuse forecasting algorithm which uses offending history data to forecast the nature of future domestic abuse, could be efficient at predicting future instances of serious abuse. In the police force analysed, around half of all serious domestic abuse arrests and more than half of the less-serious domestic abuse arrests, had prior arrest records in the preceding two years, and so could potentially be forecast by our model. Of these, the random forest model we developed could successfully forecast 77% of the serious cases and 85% of the less serious cases. When forecasts for no future domestic arrests were made, the model was almost always correct although because it was designed to favour the accuracy of the serious arrest forecasts, the model had a comparatively low efficiency rate in its serious forecasts – 90% of which would not go on to be arrested for serious abuse. However, the other side of this coin is that the model could give police the opportunity to intervene in more than a third of all serious domestic cases before they occurred.

The findings presented here also suggest that improvement on these results be possible. Extending the timeframe over which the model forecasts could increase its reach. Most arrestees for domestic abuse had some form of prior arrest, and in most cases, some form of violent arrest – just not necessarily within two years of their domestic crime. Our analysis indicates if the forecasting window were taken to five years – which is to say that an arrestee would be processed with a view to forecasting what might happen with five years of their arrest, then around two thirds of serious abuse could potentially be predicted.

In the next chapter, the implications of these findings are discussed in detail. The findings here present many interesting aspects for debate in both practical and theoretical contexts. These findings suggest that a proportion of serious domestic abuse can be forecasted successfully and questions about the efficiency and morality of the procedure then follow.

20 Discussion

20.1 Chapter roadmap

This chapter discusses the findings in the context of the existing theoretical, practical and research landscape set out in chapters 10 to 12. The aim here is to confront the important 'so what?' question for each finding and expand on existing theories explaining domestic abuse. It begins with a synopsis of the findings in relation to the research questions across the five themed areas. So far, we have seen that for most offenders and victims, abuse is a single occurrence in police records, with no evidence of escalation among those with five or more records. Although there is a high concentration of harm among a small proportion of offenders, 40% of those 'power few' have no prior, nor subsequent record of abuse. Even among those with prior records, many have only one or two cases which the police might use to accurately forecast the risk of future serious crime. To this end, we have demonstrated how arrest records for *all* crimes, might be utilised to predict future abuse. Most domestic abuse arrestees have prior arrest records of some kind, and a random forest model, forecasting two years ahead of each arrest, could potentially identify 37% of all future serious arrests and make almost no dangerous errors. The chapter begins with a comprehensive recap of these findings and then we explore the implications of these points in more detail.

To minimise repetition, discussions concerning repeat abuse, escalation and concentration of harm (the questions addressed using Dataset 1) are amalgamated and discussed separately to the serial abuse and forecasting findings. The remainder of the chapter is comprised of four distinct aspects of discussion: (1) theoretical implications – what these findings mean for our understanding of domestic abuse theory, (2) research implications – what additional research could build on these findings in future, and what the application of these methods brings, (3) policy – what these findings mean for practitioners and the formulation of domestic abuse strategy and finally, (4) limitations – how reliable the findings are and how relevant they are to researchers and practitioners outside the jurisdictions on which they are founded.

20.2 Summary of findings

The results shown in Chapter 15 show that most domestic abuse that is reported to police is an isolated occurrence, but the probability of further domestic abuse increases beyond 'more likely than not' after the third report.

The results shown in Chapter 17 strongly suggest there is no evidence in police data to support theories of universal escalating severity. These results also go further than previous analyses in respect of escalation (Bland and Ariel, 2015; Chambers-Mclellan, 2002; Barnham et al., 2017; Kerr et al., 2017) to the extent that there is evidence for a pattern of de-escalating severity in police data. This final point does not disprove escalation universally, but makes the crucial point that, if it does exist, the police do not see it in recorded events.

The vast majority of crime harm in police data is attributable to a very small proportion of victims and separately, offenders, and around four in ten of the highest-harm victims and offenders are previously and subsequently unknown to police for domestic abuse. Of those in this 'power few' group that do have multiple records, in most cases the serious crime occurred in the first three reports.

A substantial proportion of repeat victims and offenders are linked to more than one party in domestic abuse reports. The results in Chapter 16 indicate that serial domestic abuse perpetrators are a distinct cohort of suspects which make up around 10% of all offenders in a multi-year period. Serial perpetrators are marginally younger than repeat or single-time offenders and they contribute more domestic offences and 2.3 times more crime harm per person than single-time domestic abuse perpetrators, but around the same level as repeat offenders. The types of domestic offences committed by serial perpetrators are less frequently rape (2% of serial perpetrator domestic crimes were rape compared to 4% for single-time and repeats) and more frequently criminal damage and 'other' non-violent miscellaneous categories. Serial perpetrators are more commonly associated with non–domestic abuse crimes than repeat or single-time offenders: 70% of serial perpetrators had some non-domestic abuse crime record, compared to 57% of repeat offenders and 33% of single-time offenders. Serial offender non-domestic criminality is also more harmful than that committed by the other cohorts. Cross-referencing non-domestic offending patterns with each cohort establishes that generalist serial and repeat perpetrators are the most generally harmful types of domestic offender.

Chapter 19 then showed that around 80% of all-crime arrestees were not arrested for domestic abuse in the two following years. Of the 20% that were subsequently arrested for a domestic crime, their combined domestic arrests in that period amounted to 49% of all arrests for 'serious' domestic abuse crimes and 56% of all arrests for 'less-serious' domestic crimes. This meant that around half of serious and less serious domestic arrestees had no prior arrest

within two years and were therefore impossible to forecast using our particular design. However, if we were to extend the 'follow-up' period for forecasting, this exposure rate would increase, and if set at 60 months, our model would have the chance to forecast close to two thirds of serious domestic arrests and three quarters of less serious domestic arrests.

By isolating arrests in which the subject went on to be arrested for future serious domestic offences, Chapter 19 identified that these individuals were more likely to have more prior arrests generally, and more arrests for 'serious' crimes, than other arrestees. They were also typically younger at the time of their first arrest. Arrestees who went on to be arrested for *any* form of domestic crime (serious or less-serious) were more likely than others to have prior arrests for any violent crimes.

A random forest statistical forecasting model successfully predicted 77% of all future serious domestic abuse arrests where the subject had a prior arrest in the 24 months preceding their serious arrest. The model would therefore successfully identify over a third (37%) of all serious domestic abuse crime arrests in a given two-year period. The rate of cautious, 'false positive' errors to achieve this level of accuracy was very high; 89% of cases forecast as likely to be arrested for serious domestic abuse were incorrect. When the model forecasted a case to be subsequently arrested for a less-serious domestic crime, the forecast was correct more than half the time, and successfully identified 85% of all such cases. Almost three quarters of arrestees who did not go on to be subsequently arrested for a domestic crime were successfully predicted, and when a forecast of 'no arrest' was made, it was correct more than 98% of the time. Crucially, in just 0.02% of the cases forecasted for 'no arrest' were the offenders actually subsequently arrested for a serious crime.

20.3 Theoretical implications

20.3.1 Repeat abuse, escalation and concentration of Harm

The evidence presented in chapters 15, 17 and 18 challenges the pervading theory of escalating domestic abuse, which generally states the domestic abuse cases increase in severity over time (Johnson, 2006; Pagelow, 1981; Walker, 1979, 1984) The results show that the majority of cases report just once to police and offer no or limited opportunity for escalation to be identified. This certainly does not discount the possibility that domestic abuse escalates unseen (as proposed by Johnson, 2006), but there is not yet any robust evidence to support this. What we now know is that police are typically recording higher severity crimes

first, with levels of harm then generally diminishing or stabilising over time. This should inform, if not shape, future theoretical constructs regarding escalating severity in domestic abuse cases. The current construct is somewhat vague, if not entirely outdated, and the time is right for revision. Walker's analysis (1979, 1984) proposed a cyclical escalation of patterns of violence within a relationship, which set the general template for theories of escalation, but no research has since added much by way of specific detail. Richards et al., 2008 indicated that escalation was an important indicator of future serious harm, but neither they, nor others researching escalation in general (Andersen et al. 2003, ; Dutton and Kelly, 2002; Johnson, 2006), provided empirical support for this premise.

We suggest four initial aspects for revision of the theory: (1) that escalation of severity is not present in the majority of police recorded cases; (2) that escalation, if it does exist in the cases which first come to police attention as a serious crime, occurs outside of the sphere of law enforcement knowledge; (3) police knowledge of high-repeat cases may in fact play some role in de-escalating severity or (4) relationships ending and/or incarceration of offenders stop any further abuse.

The clear pattern of rising conditional probability indicates that, once established, involvement in domestic abuse becomes persistent, for both victims and offenders. This could be influenced by the nature of police involvement itself; once an individual is known as a victim or offender, they may receive more 'attention' and a natural consequence could be more records in police databases. Whatever the reason, the concentration of high levels of harm into such a small proportion of offenders and victims provides compelling support to the theory of 'power few' individuals. Targeting such groups for interventions may further Sherman's (2007) hypothesis that such groups may yield the best chance of detecting effects in criminological experimentation. Obviously, this dimension was outside of the scope of this work, but the validation of the extent of the 'power few' is an essential element of Sherman's theory – and this work supports the notion that the concept exists in victim and offender units of domestic abuse analysis.

With a majority of domestic abuse victims and offenders coming to police attention just once, focus is also required on the issue of desistance. Put simply, what stops the abuse? Or perhaps the question should be, what stops the reporting of abuse? The majority of cases we analysed were less serious and unlikely to result in custodial sentences, so it is unlikely, that this is the primary explanation. There are several well-established theories of desistance

which may explain some of the patterns in police records. For example, what are the roles of programmed potential, social context and agency (Bottoms, Shapland, Costello, Holmes and Muir, 2004)? There is logical potential in all these elements; desistance may come about through perpetrator management or victim support, through family or community culture, or through the natural maturation of a victim or offender's personal characteristics.

It is also possible, if not probable, that separation and estrangement of couples and family members plays a key role in the prevalent desistance. The role that cessation of relationships has in preventing future violence is yet to be quantified in any empirical sense. The majority of research to date has focussed on the increased risks attached to separation (see Hotton, 2001; Kaye, Stubbs and Tolmie, 2003, for example). Our results suggest that this topic might also play a more central role in prevention than previously thought.

20.3.2 Serial abuse

The establishment of the prevalence and characteristics of serial perpetrators is founded on the premise of general repeat offending in domestic abuse cases, for which there is much previous evidence. The findings presented in Chapters 15 and 16 consolidate the previous work on concentrations and crime harm of Kock (1999), Sellin (1931), Sherman (2007) and Sherman, Neyroud and Neyroud (2016), among others, in describing the widespread occurrence of repeat domestic abuse. These results show single-time offenders compose the largest group in the dataset, but one in every four suspects features more than once and this research expands on this basis to describe a distinct sub cohort of repeat offenders that are suspects in crimes against more than one victim and thus serial perpetrators.

Previous research on serial perpetrators of domestic abuse has thus far been limited to a handful of studies that have found tangible evidence of prevalence among domestic offender populations. Estimates of prevalence have ranged from 4% (Robinson, 2017) to 43% (Bocko et al., 2004) owing to substantial variation in methodological approaches. The largest sample on which previous evidence was based was fewer than 1,500 (Bocko et al., 2004). The most comprehensive finding on prevalence from previous works was Bland and Ariel's (2015) sample of around 18,000 perpetrators. Their dataset, however, included non-crime incidents in which suspect status was nominally assigned by the person recording the event. Practitioners argue that, in a non-crime incident, no party can be designated as either victim or offender/suspect, because no crime has taken place. In the context of increased scrutiny by the police inspectorate of crime recording standards, which manifests in a prima facie

approach to crime recording, many forces have seen trends in rising domestic crimes at the expense of decreasing non-crimes (ONS, 2018), and so the perspective of 'crimes only' offered by the present analysis offers a stronger basis for conclusions about prevalence. This research therefore presents, in terms of sample size, the largest study to date to explicitly examine the prevalence of serial domestic abuse offending. In the absence of comparable studies, it is difficult to assess whether the result (prevalence of 10%) is unexpected or not. It is lower than the 17.6% found in Bland and Ariel (2015), but this is unsurprising given the differences in data because non-crime data expand the scope of the dataset to include non-criminal activity, and it is logical to conclude that the inclusion of such data in this research would increase the relative size of the serial cohort.

The prevalence in Dataset 2 was also 5% lower than the prevalence found in Dataset 1. We think this is probably partly due to the shorter time horizon – two years in Dataset 2, compared to around five years in Dataset 1. It is highly probable that, as the time horizon is expanded, the number of serial offenders grows because there is a greater opportunity for the development of new relationships and therefore abuse. A prevalence estimate of 10 to 15%, caveated as dependent on time horizon, therefore seems a sensible option for the progression. This would sit broadly in line with Hester and Westmarland's 9% (2006 – but with smaller sample and shorter time horizon) and fall roughly in the middle of Robinson's (2017) estimated range of 4–20%. More research is needed here to establish the extent of serial offender prevalence in a longer time range. We suggest that examination of periods of ten years and above would be appropriate to incorporate the impact of custodial sentences and relationship development. Cross-jurisdictional analyses may also increase the prevalence

Collectively, these estimates throw into sharper relief, the extent of this phenomenon and situate the cohort as a relative 'few', not a 'many'. One question naturally follows: is this few more harmful than the average offender? Are they major contributors to the 'power few'? Here, the results indicate that generalist serial offenders are more harmful, but do not dominate the 'power few', at least in terms of the severity of domestic abuse alone. Our findings suggest that the theories which prioritise the targeting of serial perpetrators over other cohorts warrant closer examination, and that non–domestic abuse offending patterns are as important as domestic offending classifications.

Nevertheless, since there is a cohort of serial perpetrators, comprising at least 10% of domestic abuse offenders, a number of theoretical questions naturally follow. Two such

fundamental questions concern (1) whether the criminal behaviour of members of this cohort can be classified in line with existing research on batterer typologies, and (2) why some offenders strike against more than one victim and others do not. In the case of the former question, Chapter 12 already presented several points of reference with which to compare our findings. Naturally, the definitions involved in the typologies established by previous researchers have varying degrees of translatability to the dataset used in this research, but here follows a brief attempt to cross-reference each of them against these findings.

Firstly, consider Brisson's (1981) and Gondolf's (1988) three-type models (type I – 'sociopathic' abusers, who commit high levels of physical and social abuse; type II – 'antisocial' abusers, who are generally more violent but less likely to be arrested; and type III – 'typical abusers', who are generally violent but less disposed to serious violence). This typology is based on the notion of severity yet lacks a rigorous definition of what constitutes 'high level'. It does not make reference to forms of crime other than violence and abuse, and – crucially – in every typology, the abuser is defined by a repeat course of behaviour. In this last respect, serial perpetrators may be relevant, but our findings conflict with some particulars in this classification typology. Our findings show that serial perpetrators may more frequently perpetrate domestic abuse, but not necessarily more serious crime; in fact, serial perpetrators commit more *less* serious types of abuse and more *other* types of crime in general. Our conclusion therefore is that it is not possible to make effective use of the Brisson/Gondolf classification system in the context of serial perpetrators.

Chapter 4 highlight similarities between the Brisson/Gondolf typology and Johnson and Ferraro's (2000) description of five types of abusive relationship. One of these types, 'violent resistance', does not sit well with the notion of a serial perpetrator. An offender who is retaliating could do so in multiple relationships, of course, but this is relatively unlikely. 'Violent resisters' are the other side of the coin to 'intimate terrorists', the description of whom, as emotional and physical controllers of their partners, suits the popular contemporary narrative of serial perpetrators as more dangerous individuals purposefully seeking out victims. However, the notion of an intimate terrorist as a perpetrator able to exercise a pattern of controlling behaviour does not fully accord with the data period we analysed in this work. While it is certainly possible that those people whom Johnson and Ferraro would term 'intimate terrorists' may have moved from one relationship to another within the 843-day study period, it is far from guaranteed. It is more logical that the 'repeat' cohort contains a higher proportion of this kind of offender, although it cannot be discounted that a longer

exposure period may change this balance. Instead, we argue that serial perpetrators are much more likely to fit under the definition of 'common couple violence' or 'generalist borderline offenders', for the following reasons. Firstly, 'common couple violence' occurs only 'once or twice', according to Johnson and Ferraro, but it is entirely feasible that, over a period of time, a 'common couple' offender could accumulate victims as they move through relationships. As with intimate terrorists, this is perhaps more likely to occur in a longer dataset, but the lower-harm nature of 'common couple violence' tallies with the marginally lower harm propensity (on an individual crime basis – see the tendency toward criminal damage and away from rape, for example) of serial perpetrators. Secondly, the higher prevalence of all other forms of non-domestic crime correlates well with the description of the emotionally needy 'generalist-borderline' offender. These results are inconclusive on all counts against the Johnson–Ferraro typology set, but there are parallels that can be drawn, particularly with this latter type.

Holtzworth-Munro and Stuart's (1994) work consolidating 15 other batterer typologies, which itself was later heralded as the most robust typology model by Dixon and Browne (2003), bears similarities to Johnson and Ferraro's and Gondolf's typologies. However, it is easier to draw parallels between Holtzworth-Munro and Stuart's three typologies – family only, general violence and generalist – and Dataset 2 using the non-domestic crime records from the same time horizon. As we showed in Table 26, serial offenders were more than twice as likely to be 'generalist' than single-time offenders, and 1.35 times more likely to be so than repeat offenders. While this does not offer unequivocal endorsement to these typologies, we suggest that viewing the phenomenon of serial perpetration from these perspectives may offer insight in the future development of domestic abuse theory, and these should include the notion of serial perpetrators with generalist non-domestic offending records as being among the most harmful domestic offenders.

20.3.3 Forecasting

Documented 'theories' of domestic abuse forecasting are predominantly confined to expositions of risk assessment processes largely predicated on structured professional judgements (Campbell, Sharps and Glass, 2001; Dutton and Kropp, 2000; Hoyle, 2008). A more appropriate theoretical context for Chapter 19's findings is the overarching theoretical debate about whether actuarial forecasting instruments are superior to clinical instruments. This is well-worn territory, with a host of studies examining the issue dating back several decades (Ægisdottir et al., 2006; Dawes, Faust and Meehl, 1989; Dolan and Doyle, 2000;

Litwack, 2001; Meehl, 1954), with a variety of results. Our findings further contextualise this debate by contributing one of the first analyses of a machine learning based, actuarial risk assessment tool for domestic abuse.

The assessment of future risk to victims is a central tenet of the current approach to policing domestic abuse in England and Wales (Robinson, Myhill, Wire and Roberts, 2016). The main instrument used for that purpose is the DASH, a structured professional judgement–based tool which has been subject to recent review and revision by the College of Policing and has yet to be assessed for predictive validity other than in three single-force studies (Chalkley and Strang, 2018; Thornton, 2017; Turner et al, 2019). These first two of these studies, which have matching methodologies, only assessed the predictive validity of the DASH in respect of 'lethal' and 'near-miss' domestic abuse, a definition which makes up only part of the 'serious' definition used in this research and therefore resulted in much smaller sample sizes. Both studies found a high 'false negative' rate among this type of domestic abuse. In Thornton's sample of 118 cases, none of the 13 murders and only 11% of the 'near-misses' had ever been assessed as 'high risk' using the DASH. In Chalkley and Strang's 67 cases, 45 had no prior assessment of 'high risk'. The 'false positive' rates (high risk forecasts which resulted in no fatal or 'near miss' crime) were also very high – 99% in both cases.

The third study (Turner et al., 2019), analysed a much larger sample of domestic abuse crimes with a wider definition of serious harm (assault with injury and above). They also found a high rate of error – officers correctly assessed just 5.7% of serious repeats, and a corresponding maximum 'false negative' rate of 94.3%. They argued that the DASH results were little better than chance and that officers were actually underestimating the prevalence of future serious domestic crimes by 34%.

So, it seems that high error rates in domestic abuse forecasting are not unprecedented (see Thornton, 2017 for a summary) and though we must be cautious, because these are just three studies, these results are not encouraging for the DASH. The assessment, which is typically issued in every domestic abuse case involving intimate partners, and in most forces is completed at least twice for every event, is resource intensive and therefore high cost, and with police forces recording domestic abuse in increasing numbers (ONS, 2018), it is potentially unsustainable. To this end, the College of Policing's proposed revision to the DASH is currently being tested across the country, but at the time of writing, there is no

evidence concerning its predictive accuracy. As such, it is impossible to draw a precise comparison between the random forest model and the latest iteration of the DASH but the evidence on the previous version indicates it is somewhat less effective at predicting future harm than our model, which although made plenty of 'false positive errors', had a high level of accuracy (77%) in predicting future serious arrests. What does this mean for theories of risk assessment in domestic abuse? We suggest that our findings are promising enough to merit further consideration of the actuarial versus clinical debate in this field, adding to the recent conclusion of Turner et al., (2019), that actuarial processing of offender history may improve the predictive accuracy produced by the DASH.

Nonetheless, it is too early to conclude that any actuarial model based only on arrest cases could solely replace the current method of risk assessment. With only 49% of serious arrestees and 56% of less-serious arrestees having been arrested previously in the prior two years, a large proportion of future domestic abuse could not be predicted by our model to begin with. While the model might be reasonably adjusted by extending the case follow-up period, it remains the case that not all domestic abuse cases result in an arrest (ONS, 2016a; 2017). Of course, as Chapter 18 highlights, many serious cases have no prior DASH assessment either, because the serious crime is the first time the police are aware of the case – so the DASH too has greater limits on its predictive scope than our model. There is also some logic that most serious cases will involve an arrest, and if we accept this notion then the predictive accuracy of the random forest model has a superior predictive accuracy than that shown by studies of the DASH thus far – a notion supported by Turner et al (2019). The random forest model has a 63% 'false negative' rate in terms of identifying overall serious domestic abuse (including those cases without prior arrests), superior to the 'false negative rates' of the DASH found by Thornton (2017), Chalkey and Strang (2018) and Turner et al. (2019). Furthermore, if we match the conditions of this comparison and take the 'false negative' rate of the random forest model as applying only to the serious cases it could potentially identify, then the rate drops to just 23%. Furthermore, the 'false positive' rate of the random forest model is more efficient (89% compared to 99%). These comparisons are by no means perfect, but they provide promising evidence that actuarial forecasts can improve upon the current dominant theory of structured professional judgement-based instruments.

The results show that it is possible to build an actuarial model that will predict future domestic abuse with some measure of success. Even in the harshest light, a tool which could meaningfully forecast a third of all future serious domestic abuse crimes would be of use.

Further to this, the results show that the model can successfully identify most future arrests for less-serious domestic abuse, and when forecasting no future arrest, is almost always correct (99%). The question then is not 'can we do it?', but 'should we do it?' and this is an issue for policy rather than theory.

20.4 Implications for future research

20.4.1 Repeat abuse, escalation and concentration of harm

The desistance of the majority of domestic abuse offenders and victims requires further explanation. While the natural inclination may be to focus future research efforts on the 'power few' or the serial cohorts, both the high- and low-harm branches of the desisting cohort may offer insights into prevention theories. These groups provide fertile ground for understanding the role of quality of service of justice agencies in maintaining victim engagement. Police forces in England and Wales are now required by law to commission surveys of domestic abuse victims that may offer a useful potential data source, and forces may wish to consider question sets that help to contextualise desistance theories.

Surveys of communities with disproportionately low reporting levels should also play an important part in the future research landscape for understanding both 'never called before' cases and escalation. Both issues appear to require explanation involving more data sources than only police records. In relation to 'never called before' cases, there are three primary research questions to be addressed: (1) Do the police know about the victims or offenders in these cases in any other guise than domestic abuse? (2) Do any other organisations have knowledge of the victims or offenders prior to the occurrence of serious harm? Finally, (3) are 'never called before/again' victims subject to unrecorded incidents of domestic abuse? This last question is also pertinent to escalation, but studies examining the extent of unrecorded incidents need to be aimed at populations wider than just never-called-before cases, because they need to address another important question: is an escalation in severity masked by the underreporting of domestic abuse? Research in this area should focus on the specific issue of relationship cessation and estrangement of family members, which may hold valuable insight into explaining desistance, if indeed, that is what is happening in these cases.

A fourth area of further research might concern the chronic victims and offenders who remain comparatively low harm. Like the desistance cohorts, there may be useful practical

and theoretical insight in understanding what factors correlate with the absence of serious harm in spite of the repeated occurrence of abuse. Such research would also be useful in developing the understanding of common couple violence (Johnson, 1995).

The types of data that are most predictive of harm are particularly pertinent. Most crime records feature data that are linked to investigations and the immediate prevention of risk, yet research in psychology and psychiatry suggests that robust predictors of domestic abuse are less often behavioural and more often linked to the personality of the abuser. Space constrains prevent us from exhaustively reviewing this line of research (see review in Johnson and Sachmann, 2014), but interviews with batterers and those who have committed successful or attempted femicide have shown that 'the common themes identified among the participants in the interpersonal context were … their strong need for control along with excessive dependency on the intimate partner or on the relationship with her. This dependency is perceived as desperate or pathological love' (Mintz, 1980; Elisha, Idisis, Timor and Addad, 2010, p. 509). Malmquist (2007) raised the issue of psychotic depression as a factor in domestic homicide, whereas Bernard, Vera, Vera and Newman (1982) found that domestic homicide often occurs when the woman has threatened to leave the relationship. Male sexual jealousy and male sexual possessiveness are frequently cited causes of intimate femicide across cultures (Baker, Gregware and Cassidy, 1999; Polk, 1994; Wilson and Daly, 1998; Wilson, Daly and Daniele, 1995 (Goussinsky and Yassour-Borochowitz, 2012, p. 553). Notably, the items on this list of characteristics are often unrelated to the immediate role of the police officer attending a domestic violence crime scene. Nor do second responder units collate this information, which can be captured only from in-depth interviews and lengthy research assessments, which are not part and parcel of the response of these policing units. Hence, if predictors of harm are hidden from police records, more attention is needed for interagency collaboration and valid assessments of harm – all of which are presently beyond the reach of police forces (see Ariel, Weinborn and Boyle, 2015; Florence, Shepherd, Brennan and Simon, 2011; Shepherd, 1990). If founded, this implication would further underscore the critical importance of agencies other than the police in anti-domestic abuse strategies. Future research could seek to investigate the predictive power of each of these dimensions alongside police records (the random forest model presented in Chapter 19 could be a template for such research). This is, of course, easier said than done. Information-sharing legislation raises obstacles to law enforcement agencies accessing the sorts of datasets that contain these data.

20.4.2 Serial abuse

As already discussed, Chapter 16's findings present parallels with existing domestic abuse offender typologies (Gondolf, 1998; Holtzworth-Munro and Stuart, 1994; Johnson and Ferraro, 2000), and they provide the strongest sample yet for establishing the prevalence and characteristics of serial perpetrators. But they also pose more questions to be answered. Some of these questions are a product of the new evidence established by this research, while others are born of its limitations.

Having established prevalence at 10–15% of serial abuse, the next step is to seek replication. The opportunities to further explore the prevalence of serial perpetrators are plentiful, with every Home Office police force now required to identify domestic abuse offences as a matter of course. There are three principal areas for replication studies to address. Firstly, do the findings translate to metropolitan areas? While the jurisdiction subject to this analysis contains large urban conurbations, only part of dataset 1 could be described as metropolitan. Secondly, do the findings translate to different regions of the country? Our primary dataset for serial abuse analysis was drawn from just one region. Thirdly, by how much does the prevalence of serial abuse differ with extended exposure periods? Logically, the number of repeat or single-time cohort offenders 'converting' to serial offending should increase with time, but the prevalence may change in either direction depending on the size of the other cohorts. Replications over three, five or 10 years would be very beneficial to advancing knowledge of prevalence. Any replications should also carefully consider the possibility of scrutinising relationship data, which were not available in this study. It is possible that there is a meaningful distinction between intimate partner and intra-familial domestic abuse which should be factored into the consideration of prevalence.

Serial perpetrators of domestic abuse constitute a small, but high-frequency and disproportionately high-harm cohort. However, one that makes up a minority of the 'power few' and that is more generalist and more harmful than other cohorts in its non-domestic offending patterns. These findings highlight the need for more detailed research to profile who the members of this cohort really are. Such research might focus on three areas, the first of which is psychological, to build on existing knowledge of serial offenders and to expand our understanding of the subtypes developed here. In this research, it has been shown that, when analysed through the prism of prior offending, there are distinct subgroups of serial offender. The next step would be to supplement this by developing research methodologies to

understand whether there exist any differences in the psychological profile of the subtype, or indeed between this and the other cohorts.

Similarly, there is now a need for a more forensic understanding of different 'criminogenic needs' (Ward and Stewart, 2003). In offender management nomenclature, this term usually refers to areas in which an offender requires assistance or solutions to help desist in offending behaviour, such as housing, substance misuse or mental health issues. Contemporary domestic abuse perpetrator programmes such as 'Drive' and Multi-Agency Tasking and Co-ordination (MATAC) have been developed on the basis of addressing complex needs (Brooks-Hay and Burman, 2018). Improving understanding of the evidence of these needs, particularly if there are macro-level differences between cohorts, may enable models to be revised or fine-tuned to the benefit of outcome delivery. They may also improve understanding of the extent to which existing perpetrator management schemes such as Integrated Offender Management (which is popular among English and Welsh police agencies) may be suitable for, or already deal with, portions of the domestic abuse offender population. This type of offender management is under-evaluated (see Williams and Ariel, 2012) and a worthy focus for future research efforts, given the particular focus of police on perpetrator management (HMICFRS, 2017, 2019).

The serial perpetrators identified in this study were known to the police, based on police-recorded crimes. We have already discussed that it is probable that a sizeable proportion of domestic abuse does not make it to police databases. Examining partner agency information in parallel is challenging, with the definition of serial status providing one notable obstacle (Robinson, 2017), but it is entirely possible that much could be learned (not least in relation to the two aforementioned aspects of psychology and needs profile) about the known cohort of serial offenders through the exploration of other agency data sources relating to serial perpetrators drawn from police records.

Besides establishing the scope of the serial cohort and some aspects of its nature, this research has provided an additional perspective on two other cohorts of abuse perpetrators – single-time and repeat – both of which have notable differences from the serial cohort. Single-time offenders make up 75% of offenders and the sizeable part of the 'power few'. They also have the lowest proportion of generalist criminal history of the three cohorts. Repeat offenders are more prevalent than serial offenders by half, and more harmful in terms of domestic abuse, but less generalist (although considerably more generalist than single-time

offenders). Establishing this profile sets the stage for a new paradigm of research questions about these other two groups. The single-time offenders incorporate the 'never-called-before' cohort. Chapter 16 further illuminated this group by describing that 67% had no other recorded crime contacts on the database in the same period of time. Naturally, extending the time period – potentially by using the Police National Computer system as a data source, and utilising non-crime information such as domestic abuse non-crimes, safeguarding referrals, command and control records, or missing person records – could greatly enhance our understanding of this cohort, which presents a fundamental challenge to domestic abuse researchers and practitioners. By now the point is well established: a sizeable proportion of the most serious harm is committed by offenders who are unknown to agencies for domestic abuse (see also, Barnham et al., 2017, Kerr et al., 2017 and Bland and Ariel, 2015). Addressing the knowledge gaps around this cohort is essential to building prevention programmes that target the right group of people.

20.4.3 Forecasting

Studies of random forest criminal justice forecasting that preceded this one indicated a high degree of promise for the technique (Berk, 2012; Barnes and Hyatt, 2012; Berk et al., 2009; Berk, Sorenson and Barnes, 2012). This study joins this trend, indicating that it is indeed possible to use police records to predict serious domestic crimes with a high degree of accuracy. However, this research is 'laboratory-based' and exploratory in nature. No agency has yet commissioned the development of this algorithm for practical use, so the next logical step is to conduct field-based testing. As Oswald et al. (2018) pointed out, it is not possible to determine whether the use of an actuarial tool such as ours is 'necessary' or 'proportionate' in legal terms without systematically testing the results in practice. In this subsection, we set out future research questions and further expansion.

While there is merit in replicating our methodology with new datasets to test the consistency of random forest algorithms' predictive performance regarding similar or identical datasets, operational research questions are probably of higher priority in terms of advancing the professional and scholarly debate about such tools. To this point, this debate has focussed primarily on the specific questions that might predict risk (Campbell et al., 2001; Weisz and Tolman, 2000) and the nature of the instrument (Dutton and Kropp, 2000). We suggest an umbrella question to capture the essence of this issue: 'Can domestic abuse forecast algorithms be deployed in a sustainable, legitimate procedure?' A research strategy designed around this question might be multi-faceted, with different strands of research

testing areas such as (1) stakeholder perceptions of legitimacy, (2) predictive performance and (3) practical performance.

Perceptions of the legitimacy of algorithmic tools are largely unknown, and there are several potential avenues for formal analysis, as well as much potential benefit. Perhaps the three most important groups of stakeholders to test these perceptions for are (in no order) victims, professionals and offenders. Understanding victims' confidence in such tools will likely play a significant role in determining that of professionals and commissioning bodies, too. Professionals are a key group in themselves - if the operators of administrative overrides do not support the use of an instrument, it will likely not be used. Offenders' perceptions of the fairness of such tools may seem at first glance to be an odd thing to factor into research, but extensive research has shown that offender perceptions of fairness are a key component of compliance (Bottoms and Tankebe, 2012). These perceptions of legitimacy may of course have no effect at all on the operation of an algorithm in practice, but the important point here is that we do not yet know much, if anything at all, about this area, and in the absence of meaningful evidence, there is the opportunity for ill-informed policy to develop.

Establishing the predictive performance of the instrument in the field should be a primary and on-going concern for any agency using algorithms in criminal justice. The pattern established in Barnes and Hyatt (2012) and Urwin (2017) of tests on validation datasets comprised of records more recent than those used in the training set revealed a decline in accuracy. With arrest levels declining, domestic abuse crimes rising (ONS, 2018), and new domestic abuse legislation being introduced, it is always likely that optimal algorithm performance will require regular 're-tuning'. This research need not be complex. Any algorithm of this nature could be implemented on an experimental or quasi-experimental basis, but the configuration of such research needs to be closely informed by the political and operational circumstances of the agency wishing to deploy it.

Alongside predictive performance, practical performance is a major consideration and an appropriate target for future research in this area. To be successfully applied, any algorithm must be useable. The research for Chapter 19 was conducted over a number of months, following extensive data collation and statistical tests. In practice, the algorithm must be applied almost instantaneously and in a way that does not impede the application of safeguarding and justice procedures. To ensure this, the data feeding the algorithm must be gathered appropriately, with as few keystrokes as possible, and practitioners must be

consulted on the optimum place in a process for the algorithm to be applied. Mixed methods research should be deployed to analyse the impact on cost versus benefit, and operational efficiency, to complement the assessment of predictive validity.

Finally, we note that the main objective of potential revisions to our algorithm should be to increase the predictive power. This might be achieved in one or both of two ways: (1) widening the scope and (2) increasing the number of predictor variables. In the case of (1), as we have already discussed, by using arrest cases as the 'index' event for a forecast, we have a self-imposed limit to the model's scope; we can only predict future domestic abuse where the offender has some kind of prior arrest within the specified time horizon, and only around half of arrestees for serious domestic crimes had such a prior arrest in the previous two years. We might reasonably expand this in two ways: extend the time horizon or change the index event. If we expand the time horizon to, say, five years, we could predict as much as 62% of future serious arrests. However, an increasing number of domestic crime incidents do not result in arrest (ONS, 2018). Future efforts at improving the algorithm might amend the index event to the occurrence of a crime in which the perpetrator is a suspect, or a domestic non-crime incident in which they are a participant. Both of these options require careful consideration from ethical and practical perspectives. Forecasting individuals, and taking further preventative measures against them, is potentially contentious, but we argue not greatly different from current domestic abuse practices such as multi agency conferences (MARACs) which determine actions based on professional judgements (Stanley and Humphreys, 2014). From a practical perspective, the change would substantially increase the size of the dataset for the training algorithm. It would also greatly increase the number of forecasts to be generated by an agency on a daily basis.

The second way in which we might increase the predictive power of our instrument is by adding additional predictor variables. In theory, the random forest algorithm can cope with an unlimited number of these, although as with amending the index event, there are likely to be practical considerations involved. There are a number of logical predictors that may improve accuracy, including but certainly not limited to (1) whether the arrestee has ever been the subject of a MARAC or multi-agency public protection arrangement, (2) whether the subject has ever served a custodial sentence, (3) whether the subject is a 'serial' offender, (4) the nature of the relationship between victim and offender in domestic cases, and (5) whether the subject has previous offences against children. The inclusion of Police National Computer or Police National Database offending and arrest history is also an area which

conceivably could improve the accuracy of the model's forecasts. The potential in this area is extensive, and it is possible that iterations could be developed continuously, and though the model developed here which, narrowly focused though it may be, already has a reasonable degree of predictive accuracy and utility, the bottom line is that police data records alone are not currently enough to predict all future domestic abuse events.

In 2005, Jacquelyn Campbell, prominent among scholars in the field of forecasting dangerousness in domestic violence, assessed the state of the science in her discipline as immature. Almost 15 years on, our understanding of domestic abuse and the application of new statistical techniques have improved this position. In Chapter 11, we demonstrated how a machine learning technique might be used to accurately predict serious cases of domestic abuse before they happen. This analysis is exploratory, however, and although such statistical forecasting procedures are making their way into modern policing, they remain largely untested. Developments in this area need to be handled deftly, and with due regard to the burgeoning issue of legitimacy; such tools cannot succeed if they are not accepted by practitioners or the public. The following recommendations set out the main areas requiring attention in this regard.

As we have seen, forecasting instruments can work on paper, but our experience of them in practice is restricted. The next step in the development of such tools is the design and implementation of responsible, proportionate tests to determine whether they can perform adequately in the field. Such tests need to be carefully constructed in order to maintain the legitimacy and integrity of the instrument but should involve some form of randomisation or quasi-experimentation wherever possible.

Our random forest model focused on arrest cases in a two-year time horizon, which yields the potential to predict just less than half of all serious cases. As demonstrated throughout this work, a large proportion of serious cases have never been subject to prior police attention for domestic abuse. The ultimate goal should be to find a tool that overcomes this difficulty, and to achieve this, the scope of any forecasting instrument must reach beyond just the data known to the police. This will inevitably hold legal, practical and ethical challenges, but the harm caused, and cost incurred in serious domestic abuse at the very least warrant rigorous discussion of the issue.

20.4.4 Using targeting research alongside testing

In his article on the rise of evidence-based policing, Lawrence Sherman (2013) described three forms of evidence that police agencies could apply: targeting, testing and tracking – the 'Triple T' strategy. Much of the recent debate around the evidence-based policing movement concerns testing, specifically the generation of robust trials of tactics and strategies. Yet tests of high standards are often resource intensive, lengthy and expensive (Neyroud, 2016). The testing strand of the EBP movement will continue to improve, but in the other two T's of the Triple T strategy, there is wider potential for an evidence base to inform policing practice. Indeed, in this research, there are many examples of how targeting evidence, generated through the processing of 'big data', could be put to potentially good use in shaping police activity.

Perhaps the most pressing dilemma for police forces dealing with domestic crimes is how to cope with rising demand and still fulfil their objective of preventing harm. The facts established in this analysis illuminate a clear pathway to this: a very small proportion of offenders and victims contribute the most harm, and it is potentially possible to identify a large proportion of this group before the crimes happen, using information currently held by all police forces. The correct application of this evidence should enable a new regime of prioritisation to be tested by allowing for a sharper differentiation of future dangerousness in which there is harder evidence. It is an inconvenient truth that the current dangerousness assessments used by police have not been assessed for predictive validity, and it is of course still possible that such an assessment may provide a ringing endorsement. However, this does not alter the fact that the evidence presented in this analysis suggests there is at the least a complementary way of identifying cases, and that the target group of cases is small. The latter fact emphasises that the majority of domestic abuse cases are low-harm and singular occurrences and should be candidates for a 'delay' or 'do nothing' type approach. This notion is controversial. Both campaigners and HMICFRS demand a high standard of service to all domestic abuse cases, with understandable ethical and policy motivations. However, in practice, the ongoing allocation of decreasing resources to cases that would not otherwise become high harm is unjustifiable and inconsistent with a harm-reduction objective. By spending increasing time on such cases, it is inevitable that the police service and partner agencies have less time to spend on those cases they really ought to attend to, thus jeopardising their original aim. As prioritisation is an essential part of the future domestic abuse landscape, practitioners need frameworks to prioritise based on strong evidence, and

the evidence indicates that there is sufficient material in existing data resources to create such frameworks.

20.4.5 Integrating harm measurement tools

Hopefully readers will agree that analysing domestic abuse records through the lens of a harm measurement instrument has led to useful findings. There has been a recent expansion in the availability of tools for the measurement of harm, congruent with a greater general focus on vulnerability in policing. Yet at the time of writing, the only standardised use of such tools occurs in the ONS's annual dissemination of experimental crime statistics involving the Crime Severity Score tool. To fully realise the potential of harm measurement tools, more specific guidance is required on how they can be integrated into mainstream practices. To this end, we recommend focusing on three areas. Firstly, whatever instrument is used, it must find a suitable place for coercive control. As scholars such as Stark (2007) and Myhill (2018) have described, this offence encapsulates broader patterns of harmful behaviour. The minimum sentencing tariff for this offence is set at six months (182.5 days), one tenth the weight of a rape offence in the Cambridge Crime Harm Index. However, the Crime Severity Score (CSS) does not yet include this type of crime. Correcting this issue should be relatively simple for the ONS once it has collated enough data to generate an average, and it will be interesting to see what those averages are. This may feature as part of the second area we recommend as deserving of attention: enhancements to the CSS methodology. The present methodology includes gender and age disparities which are not justifiable. For example, the rape of a female is presented as 1.1 times more harmful than the rape of a male. The rape of a male aged between 13 and 16 is considered 1.6 times more harmful than the rape of a male under the age of 13, and so on. Addressing these methodological differences would enhance the usability of the tool. In lieu of these developments, the Cambridge Crime Harm Index is a viable alternative.

Finally, there is room for the development of outcome indicators based on harm indices. These would better reflect the intent of police and partner agencies in respect of domestic abuse, although care needs to be taken to avoid the pitfalls seen in police performance management in general. Creating perverse incentives through the use of targets is not advised.

20.5 Implications for policy

20.5.1 Repeat abuse, escalation and concentration of harm

As already described, in England and Wales, the Domestic Abuse, Stalking and Honour-Based Violence (DASH) form is a key component of the response to domestic abuse. For police forces, though application is not statutory, the DASH plays a significant role in granting access to further services. The results of chapter 17 raise concerns about the role of escalation questions within the risk assessment. The primary challenge is this: is it right that escalation of severity should play any role – let alone a potentially pivotal role – in the calculation of response when the evidence clearly states not only that there is no significant pattern of rising severity, but that the first reported incident of domestic abuse is often the most serious?

This challenge may be unlikely to gain favour among practitioners because there is an obvious logic to the theory of escalation, but we do not dismiss the theory of escalation outright. Instead, it offers strong empirical evidence that the pattern does not play out in official records. There are some very plausible reasons why this may be so; it may be that the response to an initial serious crime produces a de-escalating effect (through law enforcement or couple separation, for example). Alternatively, the escalation may occur away from police view. Moreover, it is important to note that this analysis has not focused on dyads, so the study of escalation does not only relate to specific relationships (though in the cases of single-record and non-serial repeat victims and offenders, the results also relate to dyads). Considering the findings on the extent of serial offending and victimisation, this is an important consideration on the basis that escalation may follow an individual rather than occur in a relationship-specific pattern.

We suggest therefore that there are several policy considerations. Foremost among these is the issue of risk assessment, where the main challenge continues to be how to assess the risk of serious harm with little or no prior knowledge of the subject? Our analysis suggests the focus of attention should be on some symbiotic, targetable cohorts: (1) serial offenders, (2) serial victims, (3) high-harm individuals who have desisted, (4) the previously known ‘power-few’ cohort, and (5) the previously unknown ‘power-few’ cohort. Developing strategies (which may be interdependent) for targeting each group could be explored as a means to harm reduction.

Although we have highlighted the not-insignificant proportion of high-harm cases that come to police attention just once, our analysis has also indicated the importance of prior domestic abuse in predicting future behaviour. Once an offender or victim becomes a 'repeat', it is more likely than not that they will be linked to a further incident. Chapter 16 identified 10% of offenders as 'serial', a group that commits more domestic crime as well as more non-domestic crime and contributes more harm overall. Yet it is not routine practice for these issues to be considered by practitioners in a systematic way. The number of times an offender or victim has been involved in domestic abuse is important, as is the number of victims an offender has committed domestic abuse against. Even without using complicated machine learning techniques, the standard integration of these relatively simple items into the suite of information available to call handlers and first responders could provide useful intelligence in predicting future behaviour.

20.5.2 Serial abuse

Chapter 16 poses a fundamental question to practitioners: what is the 'gain' in targeting serial perpetrators rather than other types of offender? At the time of undertaking this research, agencies in England and Wales were developing and reviewing a number of domestic abuse perpetrator programmes, encouraged by inspectorate findings that hardly any previously existed (HMICFRS, 2017). Some of these emerging programmes explicitly singled out and prioritised, 'serial perpetrators' (HMICFRS, 2019). Given the lack of evidence on the subject of this type of offender, particularly concerning prevalence and general characteristics, this could be considered a surprise. It is not clear what the catalyst was for the identification of 'serial perpetrators' as a core group to be targeted; it may be simple logic, based on the perceptions of serial murderers, rapists and arsonists as among the worst types of criminal, or it may have been public interest campaigning, such as that driven by Paladin, a national stalking advocacy service. Paladin, one of whose directors is one of the authors cited on serial perpetrators (Richards, 2004), published an 'overview briefing' (Paladin, 2014) identifying the lack of any framework for tracking serial stalkers and domestic abusers. The briefing proposed a national register, with the primary aim of homicide prevention, similar to the Violent and Sexual Offender Register used by law enforcement agencies in England and Wales. Regardless of the catalyst, the fact remains that large investments have been made in developing programmes to target serial perpetrators. For example, the 'Drive' project, led by the national domestic abuse charity SafeLives, was piloted focusing on 'priority' offenders, which were defined as 'high-harm' or 'serial' because 'this group carry the greatest risk of

serious harm' (driveproject.org.uk). Other approaches are characterised by the use of 'RFG' (recency, frequency, gravity – see Stark, 2016), a matrix which scores offenders based on offending history to establish programme entry thresholds.

But, what approaches programmes such as 'Drive' take is not our concern. A more pertinent question to be examined is whether serial perpetrators are the right cohort to be singling out. Do they make up a major part of the cohort posing the greatest risk? Such a question is fundamental to understanding whether targeting serial perpetrators over any other cohort is a worthwhile investment of public finance and specialist resources, and until now, no study has examined how much harm this cohort contributes.

Two points emanating from the results of this study suggest that caution is warranted. Firstly, consider prevalence. This study considers a suspect to be a perpetrator which, from a legal standpoint, is fallacious. Suspect status alone would probably not secure entry into a perpetrator programme (unless voluntarily), but even assuming it did, the total coverage of perpetrators would be 10% to 15%, a very small minority if 'serial status' were a criterion for inclusion. Second, how harmful (a proxy for 'risky') is that 10% to 15%? The answer appears to be quite harmful but not a great deal more compared to general repeat offenders. In Chapters 16 and 18, we defined the 'power few' (Sherman, 2007) as those causing offenders 80% of harm. Applying this threshold to Dataset 2, Chapter 16 showed that the most harmful group therefore comprised just 1,081 out of 17,641 perpetrators. Serial perpetrators were disproportionately represented in this small group, compared to their prevalence in the overall offender population within the dataset - they made up 10% of all offenders in Dataset 2, but in the power few, they comprise 17%. This was still well below single-time offenders, which make up 56% of the power few (compared to 75% of the whole dataset). It is important to further clarify these statistics. There are more serial offenders in the highest-harm group than an even distribution would yield, indicating that, at the most basic level, a serial perpetrator is more likely to commit a high level of harm than a single-time offender. However, 84% of the highest-harm offenders were not serial perpetrators. Repeat offenders were even more disproportionately represented among the 'power few'.

It is important to also consider the prevalence of the 'power few' within each cohort, which further emphasises that serial offenders do not have a monopoly on high-harm domestic abuse. A tenth of serial perpetrators in the dataset were in the most harmful 'power few' group – more than double the rate of single-time offenders (5%) and practically the

same as repeat offenders (11%). This means that, while serial perpetrators may be more likely than single-time offenders to commit high-harm offences, they are as likely to do so than repeat offenders. These points are important for the development of practice because they emphasise the need to understand distinctions between serial and repeat offenders, where there is at least some possibility – within police data at least – of predicting a high-harm outcome and intervening.

The second important point emerging from this research is that non-domestic offending history is an important predictor of high harm risk. Returning to the Holztworth-Munro and Stuart (1994) model of domestic offender, Table 26 displayed the mean CCHI score for each offender type, cross-referenced by cohort (single, repeat, serial). This table showed that firstly, on average, generalist offenders were generally more harmful than non-generalist ones; secondly, single-time offenders were, on average, less harmful whatever their prior offending history; and thirdly, generalist repeat and serial domestic offenders were, on average, the most harmful. Based on this evidence, it should be difficult to justify *not* examining the extent to which prior offending might predict high-harm domestic abuse. Chapter 19 showed that prior offending could be used in this way.

Table 26 also highlighted considerable disparity within the serial cohort. To date, developing programmes have only identified one kind of serial domestic abuser (the generic 'serial'), but our research suggests that there is substantial variation in the average harm within the serial perpetrator cohort itself. Only 6% of 'family only' and 6% of 'violent only' serial offenders were within the 'power few' group, compared to 9% of generalist offenders. But these findings point towards differences in probability worthy of further exploration, and worthy of consideration when deciding where to invest millions of pounds of public money in perpetrator programmes. This might be achieved by the relatively simple scanning of an offender's domestic and non-domestic offending history at the point of assignment to an offender programme, with some weighting attached to respective results (e.g. more generalist serial perpetrators might receive a more intensive programme), but it is important to stress that this research is preliminary, and more exploration of predictors is needed to supplement the simple probabilities shown here.

Then there is the question of a serial abuser register. There is no way of knowing from this work whether such an approach would be beneficial in practice, but it does indicate the extent to which a register would deal with high–harm domestic abuse. With prevalence at

10% to 15% of domestic abusers, of which more than three quarters do not become 'high harm', by the threshold used in this research, targeting serial offenders would at best be a highly imprecise endeavour. Even if all suspects were to be entered on to a register, the majority of the resources invested in policing these individuals would be spent on cases that would never progress to high-harm status. Moreover, there is no legal basis for a suspect who has never been convicted – as is the case with the majority of such offenders – to be placed on such a register (Sexual Offences Act, 2003). As such, it is highly likely that any such register would have marginal impacts at best.

20.5.3 Forecasting

There are a number of practical implications for practitioners wishing to implement an actuarial forecasting model based on prior offending record, such as our model or one similar to it, into their domestic abuse perpetrator management processes. We assess this from three perspectives set out in previous studies on actuarial tool deployment in criminal justice organisations: the political implications (Berk, 2012), the practical implications (Barnes and Hyatt, 2012; Berk, 2012; Berk et al., 2009) and the personal implications for the individual subjects of the forecasts (Berk et al., 2009).

20.5.3.1 Political implications

No algorithm will perfectly forecast all cases (Berk, 2012), so there is a decision to be made about what level of error is tolerable. In a domestic abuse context, this question is particularly sensitive. Commissioners are likely to be highly reluctant to be seen to 'accept' any form of incidence of domestic abuse, particularly serious abuse. Yet we argue that, in practice, they are doing precisely this. In the four years covered by Dataset 3, there were 1,398 arrests for serious domestic abuse, of which 681 had the potential to be forecast because the offender had been arrested (for any type of crime) within the two years preceding the arrest for the serious domestic crime. Of those 681, 523 would have been correctly forecast. The primary political dilemma facing commissioners is therefore multi-faceted. Is 523 out of 681 cases sufficient to proceed with such a model; can the remaining 23% be identified by another means of targeting? If not, is 77% an acceptable rate of successful forecast? It would certainly seem to be superior to the current rate (albeit the current instrument has hardly had rigorous evaluation), but is this enough? After all, a successful forecast does not mean that the crime will certainly be prevented; it merely presents an opportunity to do so. To achieve prevention would require financial investment in interventions for the individuals forecast as 'high risk', in the knowledge that 95% of those people with that forecast would not go on to

commit serious abuse anyway. To protect a few, commissioners would need to sanction intervening with many.

A final political consideration must involve a competent media strategy (Oswald et al., 2018). Recent examples of media coverage of algorithms have questioned the justification for such processes, focusing on the cases, places and people which are not allocated resources as a result. In relation to domestic abuse, where the police inspectorate places such emphasis on widespread robust response, and which involves highly sensitive cases with vulnerable victims and dangerous offenders, this scrutiny would likely be even more intense. Any agency wishing to implement such a tool would need to be able to secure the 'buy-in' of its partner agencies and victims as well as its practitioners.

20.5.3.2 Practical implications: the need for more information

Securing practitioner support is arguably more of a practical consideration than a political one. No domestic abuse actuarial tool could work without the consent of the operational staff who are to apply it and, while achieving this buy-in is likely to require ingredients unique to each different agency, there are probably some common principles (Strang, 2012). The factors which may contribute to securing external legitimacy – structured experimentation and administrative overrides – are also relevant to internal legitimacy. Experimentation is discussed in more detail in the next subsection, but with respect to achieving operational support, it suffices to say that the evaluation of models such as ours must be designed in a way that is transparent and rigorous and includes elements of clinical cross-examination. In her research into the Harm Assessment Risk Tool developed for Durham Police, Urwin (2017) included a strand of analysis comparing the output of the algorithm to clinical judgements on the same cases. The purpose was not to underscore conflict but to enable a more detailed understanding of differences. This kind of thoughtful research design has potential relevance for the implementation programme of any actuarial tool. Operational staff should be encouraged to 'own a stake' in the development of the instrument. This could occur at multiple stages without compromising the mathematical integrity of an algorithm. Staff may be consulted on the types of predictor values that should be considered or the data quality of those that are being considered. They should be invited to 'sense-check' the output of the forecast or undertake cross-reference exercises such as Urwin's. Most critically, though, it should be essential for them to be involved in planning for the operational execution of an algorithm, contributing to the design of any IT platform, the process map, and

the administrative override process. All such steps are likely to contribute to, though not guarantee, internal legitimacy.

20.5.3.3 A working model for prediction

In Figure 22, we show a hypothetical process map for how our actuarial tool could be integrated into a police force's existing domestic abuse risk assessment structure. Below, we briefly describe the steps.

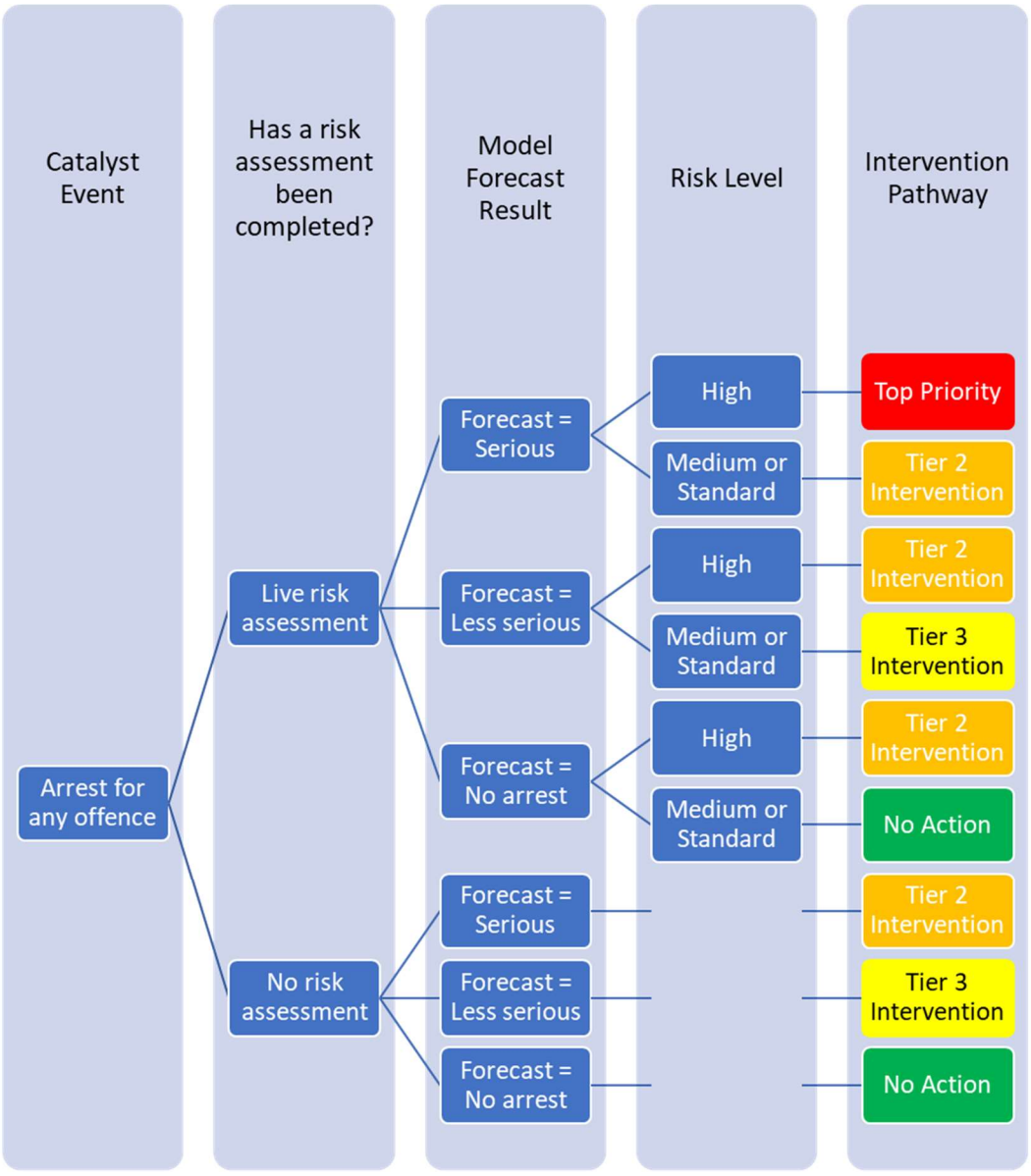

Figure 22. Potential model for the operation of a domestic abuse forecasting instrument

In this procedure, the catalyst event is an arrest (column 1). At the point at which an offender enters custody, for any offence, their case is processed using the forecasting tool. The results of the forecast then inform what offender management pathway is taken, depending on the results of a cross-reference with any existing risk assessment, as follows:

a) Cases in which an offender is forecast as having a further arrest for a serious domestic offence and which are assessed as high risk would be assigned 'top priority' status. This may include intrusive actions such as surveillance, prioritisation of intensive offender management programmes or elevated MARAC status. The details of the precise intervention are secondary in importance to the overall priority level placed on these cases, which would take precedence over all others.

b) 'Secondary level interventions' would be applied to cases that have (1) an offender forecast as having a further serious arrest but have no risk assessment or have a medium- or standard-risk assessment, or (2) a high-risk assessment and an actuarial forecast of an arrest for a less-serious domestic crime or no arrest. A secondary level intervention might consist of similar activities to those used in top priority cases but would take lower priority in triage circumstances.

c) A third level of intervention, consistent with the type of workshop implemented in the CARA experiment (Strang et al., 2017), would be applied to cases where the offender is forecast to have a less-serious arrest and (1) a risk assessment indicates either medium or standard risk, or (2) has not been completed. Although the detail of the intervention is less important than the priority, police forces wishing to reduce future demand may wish to consider proportionate investment in this category, which should produce substantially larger numbers of cases than other levels of intervention.

d) For cases forecast as having no future arrest, which either have no risk assessment or a risk assessment with a standard or medium risk level, no further action would be taken beyond business as usual.

These scenarios are hypothetical and used here only to illuminate a discussion about the personal implications of the model. While the 'devil' of full implementation lies in the 'detail' of whatever interventions were applied, we can broadly analyse the implications for

individuals that emanate from such a process (Fixsen, Blasé and Naoom, 2009). At the most fundamental level, perpetrators receiving 'serious' forecasts could potentially receive more 'service', a component of the 'risk society' (Beck, Lash and Wynne, 1992). This may be construed negatively (punitive measures) or positively (access to counselling services, peer support) depending on the nature of the intervention applied. Considering such details should be an essential component in any agency's implementation of such a model and would doubtless influence some aspects of political considerations, too. At the same time, there are personal implications here for victims as well. Intrusive enforcement, for example, instigated because of the output of a 'serious' forecast, could feasibly have impacts on the victim's personal life. It is well known that many victims do not support police investigations into domestic crimes (see Dawson and Dinovitzer, 2001; Hanna, 1996; ONS, 2018), and it seems likely these sentiments would extend to supporting perpetrator management. This may be particularly contentious in cases where an offender, arrested for a non-domestic offence, subsequently receives a secondary level intervention for domestic abuse. The complex nature of these considerations demonstrates precisely why further careful research into the operationalisation of such tools is essential. It is not simply the algorithm that has the potential for personal implications, but also the way in which it is used.

20.5.3.4 Building police forecasting capabilities

Techniques such as random forests are complex and require the supervision of experienced statisticians. Algorithms in policing risk becoming the 'shiny new thing' that can be used to reduce demand or improve performance, and this risk should be closely guarded against. Although most police forces surely employ personnel with statistical skills, the capabilities required to build and test algorithms, particularly machine learning instruments, are unlikely to be widespread yet. The danger in forging ahead with the development of such tools is that they fail and legitimacy is damaged, either among victims, practitioners or commissioners. Not all algorithms are the same, but it is not difficult to conceive the circumstances in which a small number of high profile 'failures' damage the appetite for a large number of opportunities by reducing the legitimacy of such tools. While 'buying in' the right skills could be expensive, so too could developing an in-service skill set in these kinds of statistics and their implementation. However, the latter promises the benefit of smoother customisation of algorithms, in which the authors are able to better tailor instruments to the needs of practitioners. Forward-thinking police agencies already integrate these skill sets into their recruitment plans.

20.6 Limitations

The following subsections itemise the main themes of the limitations of this work. Some of them have been discussed already but they are all worthy of repetition.

20.6.1 Police records are not the whole story

This research derives strength from the scale of the datasets it analyses; however, these data are limited to those incidents of domestic abuse recorded by the police. As widely acknowledged, and already touched upon, a large proportion of domestic abuse probably remains unreported (Carrell and Hoekstra, 2012; Gracia, 2004; ONS, 2016a, 2017, 2018), and while this gap appears to be declining, there is no immediately practicable way to address the underreporting issue in analyses like these. The fact remains that the unreported segment of the data could alter the findings and we have no meaningful way of estimating that possibility. It is also the case that some small variations in recording practices still exist between the police forces – as reflected by the HMICFRS comments on crime recording standards and the maintenance of data quality standards (HMICFRS, 2014b). The extent of data quality issues has not been estimated in this study, and this is an area that future replications may wish to address.

20.6.2 The CCHI is imperfect

The use of the CCHI is affected by the underreporting of general crime and has its own limitations (see Ashby, 2017; Sherman, Neyroud and Neyroud, 2016). In Chapter 3 we judged that the CCHI represents the best available measure of crime severity, however, at a fundamental level, it represents one specific interpretation of harm, and on a practical level, it relies on extensive manual coding of incidents. It is fair to say that the harm scores attributed to victims may be disputed by the victims themselves, and any interpretation of general statements about harm attributed to victims or offenders must consider this context. It was also impossible to control for variables that may have influenced the level of harm, such as police involvement or imprisonment of the offender. Adding such an element to the database would considerably improve any future analysis of this kind.

20.6.3 Intimate partner abuse versus domestic abuse

Police domestic abuse records in England and Wales include crimes involving family members – sibling on sibling, parent on child or vice-versa. Some agencies record details of the relationship between the victim and offender, but it is not a mandatory requirement and so data are not consistent. Adding this variable to our datasets would have been valuable

because it would have enabled us to understand any differences between 'familial' abuse and 'intimate partner' abuse in terms of distribution of harm, patterns of escalation and conditional probability. It may also be the case that a proportion of serial offenders are not moving from relationship to relationship, but from family member to family member. It was not possible to examine these issues, but we encourage replications to consider them if there is the opportunity.

20.6.4 Limited scope for prediction

While our methodology was designed to address common procedural problems, some of issues remain, particularly in respect of data reliability. As already described, Dataset 3 did not provide our model with the opportunity to predict all future domestic abuse cases, in part due to the limited time horizon for the forecasts, and in part because not all domestic abuse cases have an associated arrest. The first of these factors may be adjusted in future replications, but there is little that can be done about the second. While, in practice, this is also a reality for existing forecasting instruments, and it is a common problem that serious domestic abuse has often left no prior indication (see Chapter 18), any agency implementing a forecasting tool is likely to want to maximise its potential coverage.

We offer two suggestions to address this limitation in any future replication. Firstly, the time horizon could be reasonably extended to up to five years. This may present immediate data issues to police agencies, which have been statutorily recording domestic abuse only since 2016, but in time this problem will diminish, and for many it may not be an issue at all. Extending this time horizon could increase the maximum amount of domestic abuse arrests that *could* be forecast to closer to three quarters (see Table 33). We were able to determine this by looking retrospectively at our training dataset and counting for each domestic abuse arrest, the extent to which the offender had prior arrest records in the preceding years. Secondly, we recommend the testing of models which adjust the index event from arrest to crime report. This may be more difficult to realise in practice but this is the central point of this recommendation: by changing the point at which a forecast is run from an arrest (in the current model) to the reporting of a crime in which the subject is a named suspect, the model could increase exposure to pick up subjects who were not arrested, and thus be more resilient in the context of falling arrest rates.

20.6.5 Data quality

In an attempt to satisfy Gotfreddson and Moriarty's (2006) points regarding data dynamism, ethical considerations and data reliability, the predictor variables we selected were all taken from a police database. Most predictor variables relate to a subject's previous offending history, with distinctions between the types of offences providing some of our framework. As such, the correct classification of these data is an essential component of the process, but entirely out of our control. Police data classification has undergone much scrutiny and criticism in recent years (see Ariel and Bland, 2019), and with domestic abuse 'flagging' not required by the Home Office until 2016, it is possible that data quality problems predating this may have affected forecasting performance. These are real problems that police forces must deal with in everyday practice. For this research, a limited programme of data cleaning was undertaken, but it was not possible to scrutinise tens of thousands of records on a 'case-reading' basis. Our recommendation for addressing this issue is generalist; when conducting any future replications, researchers need to conduct a data quality assessment at the outset of their work. It is important to consider the outcome variable as well as the predictor variables, and it is equally important that any associated data-cleaning process does not transform the data beyond a state that it is practicable to achieve in a real-time environment. If this occurs, it is unlikely that the resulting model will take root in an organisation without substantial IT investment in dynamic data transformation, regardless of how good its predictive accuracy may be.

Furthermore, we cannot ignore legacy issues with crime recording practices and infrastructure. To build large datasets we had to explore some data pre-dating the relatively recent drive to improve crime recording practices in the police service. We have no measure for it, but it is probable that some of the records which would now be recorded as crimes were not recorded as such in the records we collected. Unfortunately, there is no practical option for the benchmarking the extent of this potential issue. There is only one remedy for this in future research – use more recent data – which would be inevitable in any case.

20.6.6 The black box nature of random forests

Random forest algorithms are often criticised for their 'black box' nature (Liberty, 2019), which refers to the statistical complexity of the algorithm obscuring the precise means by which a decision was arrived. For critics like Liberty, this makes it a particularly unsuitable method for application in criminal justice settings, where transparency is a key component of legitimacy. As a random forest model, ours is no different from predecessors which have

been criticised in this regard. Were it in active operation in a custody suite, issuing forecasts for future domestic abuse arrests, it would be difficult to give the subjects of those forecasts' meaningful information about the precise working of 'why' the model arrived at the judgement it did. However, this does not mean it is an impossible task. Any case that were exposed to our model could be retrieved, with the results of each decision tree conveyed, alongside an explanation of variable importance. The problem occurs in conveying this information quickly and in a manner that is easy to understand. Any random forest model is, by definition, an instrument that involves complex statistical functions. While a detailed technical report may materially answer questions, it is likely to require the interpretation of an expert to do so. Our response to this criticism is that it can be equally difficult – if not more so, on occasion – to explain the logic of a human decision-maker in a way that is as 'evidenced' as a model such as ours. This is a philosophical rather than a practical point, and ultimately there is little escape from the 'black box' nature of the model other than the emphasis on the possibility, albeit impractical, of explanation on a case-by-case basis.

20.7 Summary

There is a compelling case for revisions to theories of universal escalation. These revisions should not dismiss the notion entirely but should accept that in the majority of cases, the police will not be able to detect escalation because of any or all of (1) desistance, (2) because crimes simply do not follow an escalating pattern or (3) because police activity in itself prevents escalation. Desistance requires further research exploration because it is a dominant feature of police records and it might be a good or bad news story, but either way, it is still unexplained. The majority of victims and offenders feature just once in police records. This phenomenon also indicates that police were unaware of many serious domestic cases before they occurred – a major obstacle to preventative efforts. Additional research is required to determine if these cases truly are unknown to police, and if so, whether partner agencies have knowledge. The analysis presented in Chapters 8 and 11 in particular, suggests that a high proportion of offenders have some kind of non-domestic offending history. Work is also needed to understand the demographic profile of these cases to determine whether it is substantially different to the repeat reporting population.

Nevertheless, while the findings suggest that importance of 'never called before' offenders may be higher to harm reduction strategies than previously thought, serial perpetrators may not be as universally dangerous as they have been theorised. Dataset 2

showed alignment with existing batterer typologies and the evidence suggests in particular that serial perpetrators are more prone to more harmful, non-domestic criminality than other types of domestic offender. But while they contribute to 17% of domestic abuse harm however, they are not the dominant group and targeting serial offenders at the expense of repeat or single-time offenders is unlikely to make a significant difference overall. Our findings do require replication though, especially to determine if they apply to other kinds of jurisdiction, particularly metropolitan, and whether extending the time horizon substantially alters prevalence. In the absence of other evidence, our findings should prompt careful consideration about the proportionality of investment in offender programmes that prioritise serial perpetrators at the expense of repeat and single-time offenders. There is no simple answer to the dilemma of which types of offender to target to have the biggest effect on harm reduction, but our evidence clearly indicates a balanced approach would be best.

The most harmful domestic offenders are comparatively few, but using a random forest model, they can be forecast with a high rate of accuracy and a low rate of dangerous error if one is open to tolerating a high level of 'false positive' errors. In practice this would mean a high number of 'high risk' forecasts, most of which would not have turned out to have a future arrest for serious domestic abuse. But within this group, over a third of all serious arrests could be predicted successfully. This rate might go even higher, if the range of the forecast was increased from two years. In this sense, the model capitalises on the fact that most domestic abuse arrestees have some form of prior arrest record, albeit it not necessarily for a domestic crime. Crucially, these forecasts would almost result in 'very dangerous' errors – forecasts of no abuse for an offender who actually went on to commit a serious domestic crime.

These results challenge the dominant theory that structured professional judgement is the best form of risk assessment available and should prompt operational trials of actuarial models. These trials should assess the sustainability of performance, the practicality of delivery in the field and the legitimacy of such tools among practitioners and the community. The best route for these is proportionate experimentation and part of this, the model's performance may be improved by the addition of new variables. In parallel, careful examination of the political and personal ramifications of actuarial tool use is required to ensure that these promising yet complex techniques survive first contact with the real world of police operations.

21 Conclusions

21.1 Chapter roadmap

This short final chapter revisits the major themes set out so far and evaluates them against the original purpose of this research. The first subsection summarises the state of domestic abuse in England and Wales – a highly demanding and harmful form of crime, with a resource-intensive approach predicated on a range of strategies that have a pervading absence of empirical evidence. The subsection also summarises our source of data, police records and the original purpose of our analysis. The chapter then expands on a central theme emerging from our findings: although illuminating and challenging to existing thinking, police records alone are not sufficient to predict *all* domestic abuse cases. Be this as it may, our analysis demonstrates that there is much potential utility to be gained from the application of statistical techniques to police domestic abuse databases.

21.2 Original objectives

As the primary agency dealing with domestic abuse in England and Wales, the police service collects a large reservoir of information about cases, victims and perpetrators. Yet this reservoir is largely untapped in a strategic sense. The current response to policing domestic abuse is wide-ranging, complex and resource intensive. It is largely characterised by a 'one-size-fits-all' minimum standard ethos, with progression in resource prioritisation based on risk assessment. It is the norm for police officers to attend every domestic call-out, complete a risk assessment in at least every intimate partner case, re-assess that risk using a specialist, and arrest where there is the power to do so. The focus of the current policing strategy is predominantly on the response to, or preventative efforts in respect of, *known* cases. By contrast, there is little to no targeted investment in the discovery of individual *unknown* cases, and perpetrator management has been highlighted as a weakness (HMICFRS, 2016). It is not difficult to understand why the strategy is this way, given the escalating level of domestic abuse demand that the police service has experienced in recent years (ONS, 2018). This rise has meant that the 'one-size-fits-all' approach has become burdensome for police forces experiencing rising demand across other areas of their business too, particularly at a time when their resources have generally decreased (ONS, 2019). It is perhaps more surprising that the police's own data have not been leveraged more deeply in pursuit of greater efficiency and effectiveness, although it should be acknowledged that statutory regulation of

domestic abuse records did not come into effect until 2016. It has been argued that police records are insufficient as samples of domestic abuse (Johnson, 2008; Myhill, 2018), but it is not clear that population surveys offer more meaningful sources of analysis (Ariel and Bland, 2019). It is indisputable, though, that the scale of police records of domestic abuse is expanding greatly and offers the single largest source of domestic abuse data available.

This research set out to exploit the opportunity that police records present to domestic abuse strategy-makers. While the ONS national statistical summary (ONS, 2018) provides a useful overview of prevalence and justice outcomes on a force-by-force basis, it has limited use for practitioners concerned with the day-to-day management of domestic abuse strategies or scholars interested in determining the specifics of contemporary domestic abuse. Police records potentially offer value to both these interests, and it was in light of this promise that this research was designed. The primary areas of interest were to establish facts about patterns of harm, to assess the extent of serial perpetration in particular, and to evaluate the potential of police data for forecasting purposes.

Each of these goals was guided by current practical and scholarly contexts. Policing strategies commonly lack meaningful outcome measures, being constantly trapped in a cycle of measuring crime rates and investigative performance. Exploration of harm measurement has been preliminarily explored in a number of studies with promising results (Bland and Ariel, 2015; Barnham et al., 2017; Kerr et al., 2017), and here the intention was to expand these analyses to multiple force jurisdictions of varying types.

The management of perpetrators has become a particular subject of interest for the police inspectorate (HMICFRS, 2016), with an imperative now placed on police forces to seek provisions for managing offenders, particularly serial offenders. Nevertheless, before this study, the evidence on the prevalence of serial perpetrators was extremely limited. Chapter 16 was designed to fill this gap and presented the largest analysis of domestic abuse perpetrators conducted to date.

The prediction of future harm has been a major concern for domestic abuse practitioners for more than a decade, and it remains an area of heavy investment. The current methodology remains an exercise in structured professional judgement, untested and unequally applied (Robinson et al., 2016). Actuarial tools for forecasting future domestic abuse have been largely untested despite a body of evidence that such instruments might be more effective. Chapter 19 tested such a model.

21.3 More data is needed in this fight

As is the case in the fight against diseases, data about cases are some of the most valuable tools in our arsenal in the war against domestic abuse. Our analyses highlight the promise of police domestic abuse data to generate valuable insights. Using relatively simple techniques we have been able to establish the extent to which police are unsighted on serious cases before (and after) they happen. Equally we have placed in sharp contrast, just how few serious cases there are and just how little evidence of escalation there is. Our data have also provided us with a clearer view of perpetrators. We found that serial perpetrators were not responsible for the majority of high harm abuse but that those with more generalist offending backgrounds were more likely to fit this bill. In general, it appears that non-domestic abuse offending records could be a valuable predictor of future risk, and by using a not-so-simple statistical technique we have shown how that source of information might be used to give police an opportunity to intervene in high harm cases before they occur.

But equally throughout we have seen the limitations of this source of data. The true prevalence of repeat and serial cases, of escalation and the extent of the 'power few' may be masked by underreporting. While the extent of this remains unknown (and the point still remains: this evidence is generated only from what the police *know*), our findings are restricted by data quality issues. Such problems are well known among researchers familiar with police data but are fixable, and it is potentially highly advantageous to do so. The development of a statutory platform for domestic abuse data collection provides an ideal opportunity for the future improvement of data quality. The options are extensive, so some care must be taken regarding what improvements to prescribe, because collation may be resource intensive for police forces. An important area of focus should be the consistent identification of victims and offenders across force boundaries. Little is known about the extent to which domestic abuse victims and offenders migrate between police areas, or between victim and offender status, and a consistent form of identification would enable a national overview of these issues and a definitive view of repeat and serial abuse to be formed.

Elsewhere, flagging information may assist with the future enhancement of forecasting. The flagging of information whereby suspects receive flags for warning statuses such as 'mental health', 'suicide' or 'weapons', is not regulated by any standards and cannot be guaranteed as meaningful from one agency to another. However, if their consistency of use could be improved, all of these (and similar) variables may improve the predictive

validity of tools aimed at identifying future cases. Similarly, if it is logical that there is value in different police agencies sharing information in the pursuit of preventing domestic abuse, it follows that there is likely benefit in routine cross-agency data sharing. This already happens in individual cases; it is one of the core purposes of the MARAC procedure, but this deals only with known cases. There is currently no systematic information sharing among agencies for the purpose of identifying at-risk cases that may not yet have come to police attention. For certain, this is a complex field, both legally and practically. Public agencies carefully guard personal data, and the systematic aggregation of datasets from different agencies is rarely straightforward, but all agencies with roles in responding to domestic abuse – the police, hospitals, GPs, schools, charities and housing associations – may potentially derive benefits from mining their combined data, both in terms of reducing demand and possibly delivering better services to the public.

To overcome legal obstacles, inter-agency information sharing of this nature requires a specific purpose. In this case, we suggest that purpose to be, initially, exploratory analysis through the replication of the random forest method we applied in Chapter 19. Such an exercise would likely involve extensive data cleaning to match names, but this could be largely overcome through the use of algorithms such as Soundex. The predictive performance of models such as that presented in Chapter 19 may provide a baseline for the performance of expanded models that screen wider populations based on a different set of index events. Complex and ambitious though this avenue may be, the promise of big data analyses to assist professionals in delivering services for domestic abuse cases is undeniable. Evolving these techniques will require skill and care in order to maintain their integrity, but they offer perhaps the most promising line of enquiry for decision-makers and practitioners concerned with securing better outcomes for domestic abuse victims.

21.4 Final conclusions

Police records of domestic abuse are certainly not perfect. However, using them in an aggregated manor does provide a range of new insights into domestic abuse in England and Wales. Partial though the data may be, these insights give us perspective on what the police service is dealing with. The challenge of identifying harmful cases before they occur is daunting, particularly given the total volume of domestic abuse reports compared with the relatively small volume we might label as ‘high harm’. But to this end, we have demonstrated how police data may be used to refine strategies and proportionally target more resources on

the small number of individuals involved in such cases. The data may even be used to predict many serious domestic crimes before they happen. There is certainly room for refinement, but the promise of police records as a weapon in the fight against domestic abuse is abundantly evident.

Domestic abuse is a deeply personal crime with terrible consequences. Its occurrence is a blight on modern society, and the professionals and volunteers who work every day to respond to and prevent it, are deeply committed and courageous. There is immense value in the stories gathered from individual cases - these stories articulate the harm caused by offenders in a clear and impactive way, raising the profile of the cause and winning 'hearts and minds'. It is perhaps difficult for 'numbers on a spreadsheet' to convey the same impact, but we hope that throughout the course of this research, we have demonstrated that anonymised, aggregated data can also tell a compelling story - one which can provide valuable reinforcements for everyone engaged in the fight against domestic abuse.

21.5 Summary of research questions and answers

Table 36: Complete summary of findings

Question Number	Question	Findings
1	What is the prevalence and extent of repeat victimisation of domestic abuse?	Three quarters of these victims reported just once in a multi-year period; 25% of victims were repeats, and just 5% of victims reported three or more domestic abuse events. Although some cases were extremely chronic – reporting in double figures, such victims were rare (0.3%).
2	What is the conditional probability of further domestic abuse associated with each consecutive victimisation?	Probability of future crime generally increases with additional reports. As the repeat rates for victims was 25%, so was the probability that a victim presented a second time. After this the probability rose steadily, to around a 50% chance of a third event after the second, and an over 80% chance of additional calls after the twelfth
3	What is the prevalence and extent of repeat offending of domestic abuse?	The prevalence of repeat abuse for offenders was similar to victims, with the exception of the proportion of individuals with 15 or more events attributed, which was slightly higher at 0.4% of offenders compared to 0.3% of victims.
4	What is the conditional probability of further domestic abuse associated with each consecutive offence?	As with victims, the probability of offenders being linked to further domestic abuse crimes generally increased with each additional crime. Although there was a slight decline between offences 4 and 5, there is no obvious reason, and the increasing pattern was re-established by offence 6.
5	What is the prevalence and extent of serial abuse among victims of domestic abuse?	In dataset 1 the prevalence of serial victims was 13%.
6	What is the prevalence and extent of serial abuse among offenders of domestic abuse?	In dataset 1 the prevalence of serial offenders was 15%. In dataset 2, which covered a shorter period of time, the prevalence was 10%. We estimate the likely range to be between 10% and 20%.

7	Are serial perpetrators demographically different from repeat offenders[23] or single-time offenders?	There were only minor differences between serial and non-serial perpetrators. The mean age of serial offenders (n = 1,770) was a year below that of single-time offenders (n = 13,261) to a statistically significant level indicating that serial perpetrators tend to be younger than single-time perpetrators, but not younger than repeat offenders. Although females were more frequently single-time perpetrators, there was no significant difference in the proportions detected by a test for proportions. The rate of non-white British perpetrators was the same in the serial and single cohorts, but higher in the serial cohort than the repeat cohort by a ratio of 1.38:1, but again with no significance found by a t-test for proportions.
8	What types of domestic abuse crime do serial perpetrators commit and how harmful are they?	Serial perpetrators accounted for 21% of domestic crimes. In comparison, 26% of crimes were attributed to repeat offenders (who accounted for 15% of all offenders) and 53% of crimes were attributed to single-time offenders (who were 75% of all offenders). We would not expect an equal distribution, as single-time offenders by definition have at least one less crime than every repeat or serial offender. This pattern does not, however, hold for all crime classifications. Repeat offenders contributed almost twice as many rape crimes as serial offenders (28% to 15%), whereas 28% of domestic abuse criminal damage offences were attributed to serial offenders – 3% more than to repeat offenders.
9	Do serial offenders cause more domestic abuse harm than repeat or single-time domestic offenders?	55% of harm was attributed to single-time perpetrators, but this perspective needs to be considered alongside the number of individual perpetrators in each cohort. An analysis of mean harm clearly indicates that serial offenders of domestic abuse

[23] Repeat offenders in this sense are those which offend multiple times against just one victim

		accounted for more than twice as much harm per offender than single-time offenders (which makes sense because, by definition, there will be at least twice as many crimes attributed to each serial perpetrator compared to single-time perpetrator), but slightly less harm than repeat offenders. This picture is consolidated by an analysis of the 'power few'. Among the 1,081 perpetrators in this group in dataset 2, 17% were classified as 'serial', compared to 10% in the database overall. Repeat offenders, which made up 15% of all offenders, composed 27% of the 'power few', and the remaining 56% were single-time offenders. Overall, repeat or serial offenders were twice as likely as single-time offenders to form part of the 'power few' cohort.
10	To what extent do domestic abuse serial perpetrators commit other forms of crime, and how does this compare with repeat or single-time domestic offenders?	In every category of non-domestic crime, a greater proportion of serial perpetrators were linked to offending. In total, 70% of serial perpetrators (1,233 of the 1,770) were linked to non–domestic abuse crimes, compared to 57% of repeat offenders and 33% of single-time offenders, suggesting a greater tendency toward generalist offending among the serial cohort. These trends also extended to the measurement of harm. Serial perpetrators were more likely to be linked to higher-harm non-domestic offences than non-serial offenders, to a statistically significant level.
11	Is there evidence of escalating harm in each consecutive domestic victimisation?	For victims there was a general downward trajectory of the average CCHI score. A one-way ANOVA test determined the presence of statistically significant predictor for somewhere in the sequence meaning that at least one point, the result was not due to chance alone. A post-hoc test indicated that the harm in the first incident was in fact statistically significantly higher than all others in the sequence, except the ninth. This contradicts the notion of escalation of severity after the first crime report.
12	Is there evidence of escalating harm in each consecutive domestic offence committed?	The pattern was repeated for offenders except that the first offence was significantly more harmful than every other crime in the sequence.

13	What is the extent of concentration of harm among victims of domestic abuse?	Harm was highly concentrated among victims. Fewer than 3% of victims accounted for 80% of cumulative harm.
14	What is the extent of concentration of harm among offenders of domestic abuse?	This pattern was also observed for offenders. For both victim and offenders, the 'power few' were older and less frequently male than those outside the 'power few' with statistically significant results from t-tests of means and proportions.
15	To what extent do the police have prior knowledge of the group of victims suffering the most harm?	Of the 4,605 victims who were attributed to events accounting for 80% of all the domestic abuse harm, 1,904 (41%) featured just once in the dataset, indicating no record of domestic abuse before or after the serious crime. Although the 59% with multiple calls represents a higher level of repeat victimisation among the 'power few' victims than the general victim population, 41% still represents a significant proportion of serious cases in which police had no prior domestic opportunity to intervene. To compound this further, among the repeat cases in the 'power few' cohort, there was limited opportunity to forecast and prevent harm – 78% of the cohort (including the single-time victims) had fewer than five crimes in the dataset.
16	To what extent do the police have prior knowledge of the group of offenders committing the most harm?	Of the 4,349 offenders linked to 80% of harm, 1,710 (39%) appeared in just one record. The majority of high harm offenders – 73% - had fewer than five crimes. The actual window of forecasting is probably even more limited, because it is improbable that the 'serious' offence – the one which 'qualified' the offender for 'power few' status, was the last one in every sequence. Indeed, of the 7,041 crimes in the dataset which had a CCHI score of more than 548 (the nominal 'power few' threshold outlined in the previous section), 67% and 69% were the earliest recorded crime for victims and offenders respectively. Just 8% of serious crimes occurred later than the fourth sequential crime both victims and offenders
17	What proportion of all arrestees go on to commit domestic abuse within two years?	20% of arrestees (for any form of crime) were subsequently arrested for domestic abuse within two years of their arrest.

18	What proportion of serious domestic abuse arrestees have prior records for domestic abuse?	Arrestees who went on to commit serious domestic abuse had a prior arrest record *more often* than arrestees who later committed less-serious offences (Odds ratio (OR) = 2.3 (95% confidence interval (CI) = 1.6-3.1) or no abuse (OR = 3.8, CI = 2.8-5.4). Subsequent domestic arrestees (both serious and less-serious) more commonly had prior arrests for domestic abuse than those who committed no further abuse (Serious OR = 3.5, CI = 3.0 – 4.1; Less-serious OR = 2.9, CI = 2.8 – 3.1). Less-serious domestic arrestees more commonly had a prior record for a serious crime than those going on to commit no domestic abuse (OR = 1.2, CI = 1.2 – 1.3), and subsequent arrestees for serious domestic abuse had such a prior record even more commonly (OR = 2.5, CI = 2.2 – 3.0). Strikingly, almost nine in every 10 arrestees who went on to be arrested for domestic abuse within two years had some kind of prior arrest record for a violent crime.

19	Can antecedent inputs predict future serious domestic abuse cases to a high degree of accuracy?	In the police force analysed, around half of all serious domestic abuse arrests and more than half of the less-serious domestic abuse arrests, had prior arrest records in the preceding two years, and so could potentially be forecast by our model. Of these, the random forest model we developed could successfully forecast 77% of the serious cases and 85% of the less serious cases. When forecasts for no future domestic arrests were made, the model was almost always correct although because it was designed to favour the accuracy of the serious arrest forecasts, the model had a comparatively low efficiency rate in its serious forecasts – 90% of which would not go on to be arrested for serious abuse.
20	If so, which inputs have the greatest impact on accuracy?	Every predictor variable contributed something to accuracy but those which indicated previous domestic offending had the greatest influence. Age at first domestic arrest' was consistently the most influential predictor variable in terms of model accuracy – 1.6 times more so than the next highest influencer. Predictors relating to whether the presenting arrest (on which the forecast was based) related to a domestic crime, and the number of years since the arrestee was last arrested for a domestic crime were also highly ranked.

22 Appendix A: Technical Information Relating to Random Forest Modelling

22.1 About this Appendix

This appendix shows two elements relating to the random forest model described in chapters 14 and 19. The first element concerns the tuning of the model. The second element is the full detail of the 'partial response plots' of the five most influential variables described in Chapter 19.

22.2 Model tuning

The random forest algorithm features a number of variables which can be adjusted to 'tune' the model to refine overall accuracy and the balance of error ratios. Part of the tuning process takes place with reference to the 'out-of-bag' (OOB) error rate (Breiman, 2001), which is created by testing each decision tree in the forest against a segregated, randomly selected sample of the training dataset. Each tree has a different OOB sample which contributes to the overall OOB error rate. In the model construction process, we then incrementally adjust various tuning parameters to seek the optimum level of overall forecasting accuracy. This is different from 'data dredging' or 'fishing' which are terms that describe researchers looking for a significant effect when there is none. In a random forest model, we are concerned with the best possible forecast (within the parameters explained below). Table 37 shows the tuning parameters that were adjusted to optimise the model.

Table 37: Random forest tuning parameters

Parameter	Description	Input value
Split variables	The number of variables to be randomly sampled at each decision tree split	6
Trees	The number of decision trees to be constructed, each of which makes a separate forecast, the most frequent of which 'wins' and dictates the overall forecast	501
Sample size	The relative number of each outcome (no arrest, arrest of less serious abuse or arrest for serious abuse) to be selected at random for each tree.	400 for each sample

To determine the optimum number of variables used to 'grow' each tree (see Chapter 14 for the explanation on how forests are developed), model calculations were repeated with incremental adjustments for the number of variables used, ranging from three (the default setting in the randomForest package) and 10. The ceiling was set at 10 because as Figure 23 shows, the mean error rate for DV2 (future serious domestic abuse arrest) forecasts was climbing at this point and all other error rates had plateaued. Consequently, six variables were selected as being the optimal point for DV2 errors.

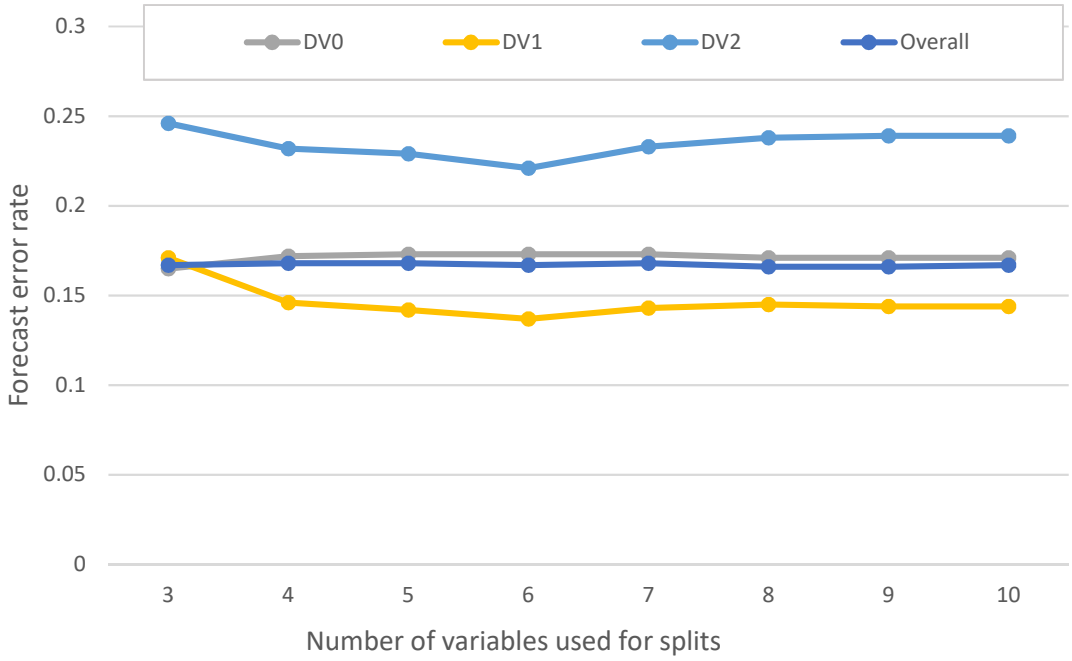

Figure 23. Mean forecasting error for different numbers of splitting variables

To find the optimal number of trees, we used the native plotting functions in the R package 'randomForests'. Figure 24 shows the cumulative mean forecasting error (based on the out of bag sampling) for each consecutive tree up to 501. The plot shows four lines, one for each outcome error rate and one for the overall error rate.

To obtain the optimal sample sizes the model was repeatedly calculated through numerous iterations of varying sample sizes. The samples were stratified by the outcome classifications and iterations processed by adjusting each in turn by 100 cases, beginning at 100 and ending at 60% of the total sample size. In total more than 70 iterations of the model were processed. The sample balance of 400 cases from each outcome classification was selected on the basis of it recording the lowest mean forecasting error for DV2 outcomes.

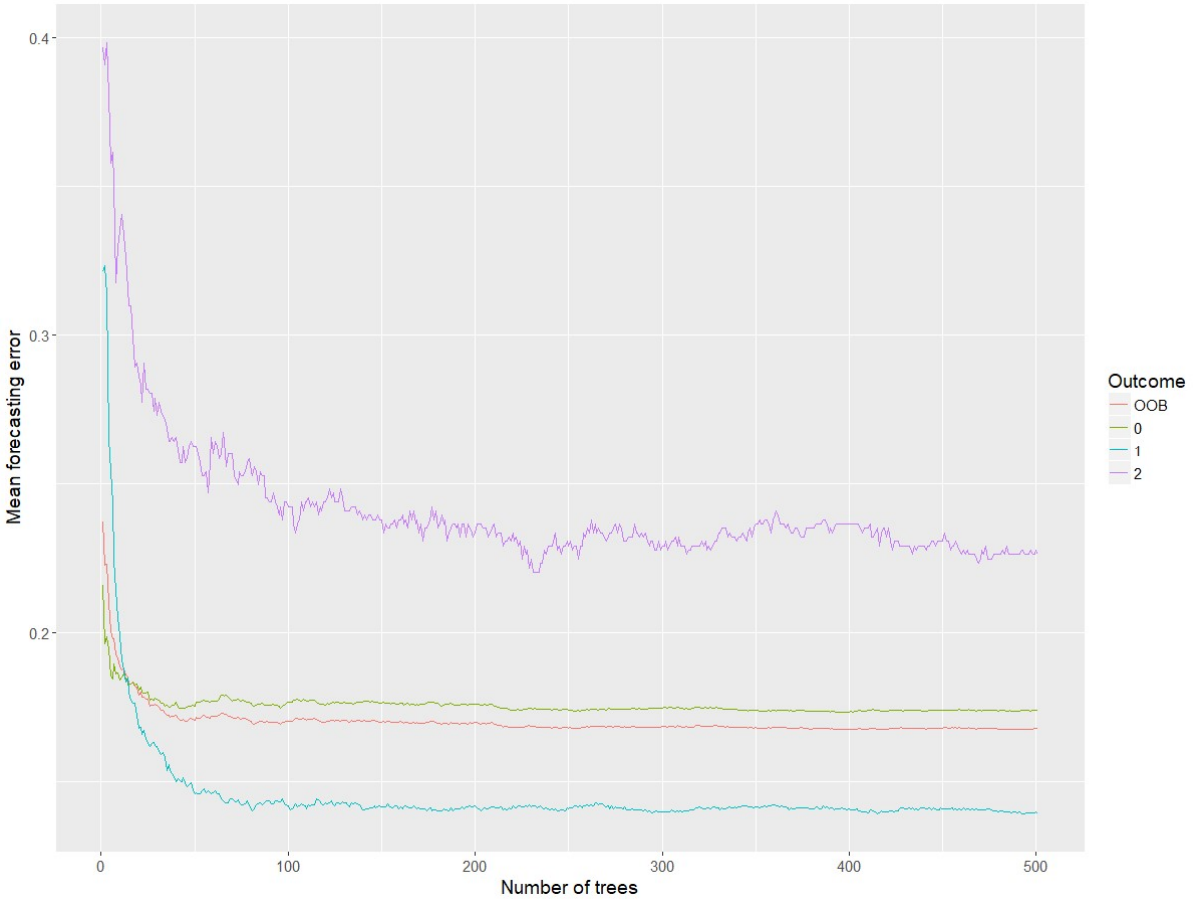

Figure 24. Mean forecasting error for random forest model trees 1 – 501

As our priority was to have the lowest possible error rate for serious domestic cases, the purple line (2) is the one we are most interested in. The other three show the greatest level of stability, with negligible difference in accuracy from around tree 100. DV2 outcomes varied somewhat more, however. The lowest error rate was around 240 trees, after which the error rate increased until after 400 trees. The comparative difference between the average forecasting error for 501 and 240 trees was less than 1%. The only real drawback to using 501 trees instead of 240 is the drain on computer processing power, but this was not considered as an important factor given the relatively small number of records in our data compared to the capacity of our hardware, so we proceeded with 501 trees. Given the small range in differences in all outcomes from trees 400 onwards we did not proceed with testing beyond this number.

22.1 Partial response plots

Section 19.5 of Chapter 19 discussed the relative influences of individual variables on model accuracy and referred to 'partial response plots' that were used to analyse this outcome. The actual plots were excluded from the section and are reproduced here in Figures 25 to 29.

Each plot has three panels, each showing the relationship between the named variable and each of the three outcomes. The 'y' axis in the plots ("yhat") refers to the centred logit of the probability that this forecast will be the outcome. The 'x' axis always refers to the unit of the variable itself.

Readers should note that negative values originate either from coding (-1 was used as the nominal value for missing/absent variables – e.g. in the case of age at which first arrested for a domestic crime, -1 would relate to no arrest) or incorrect data (e.g. incorrect dates indicating negative values in 'years since' variables).

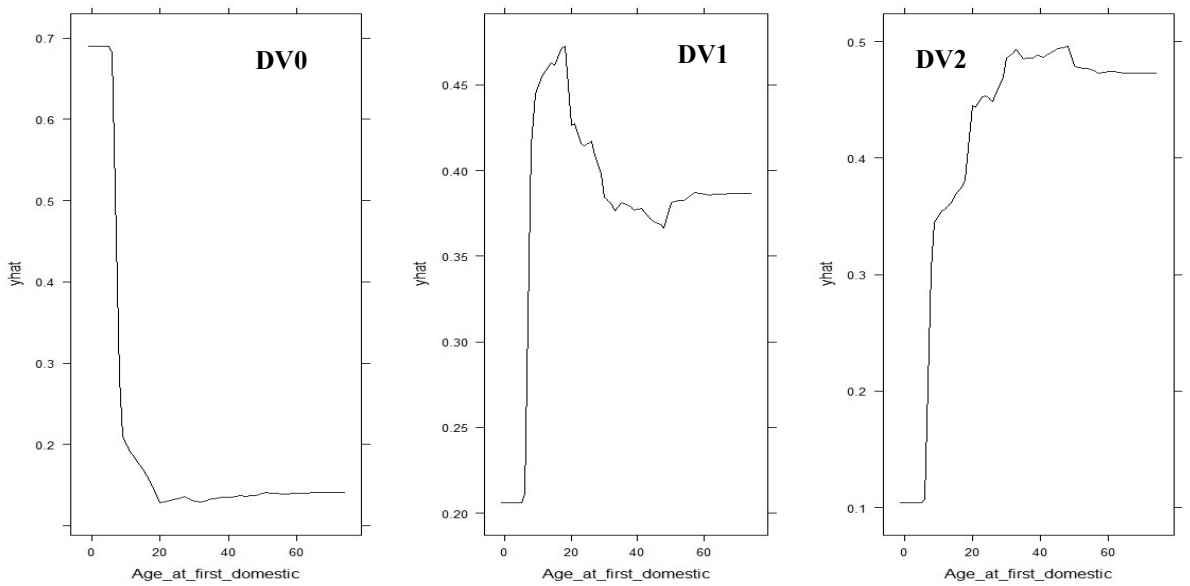

Figure 25. Partial response plots for age at which first arrested for a domestic crime

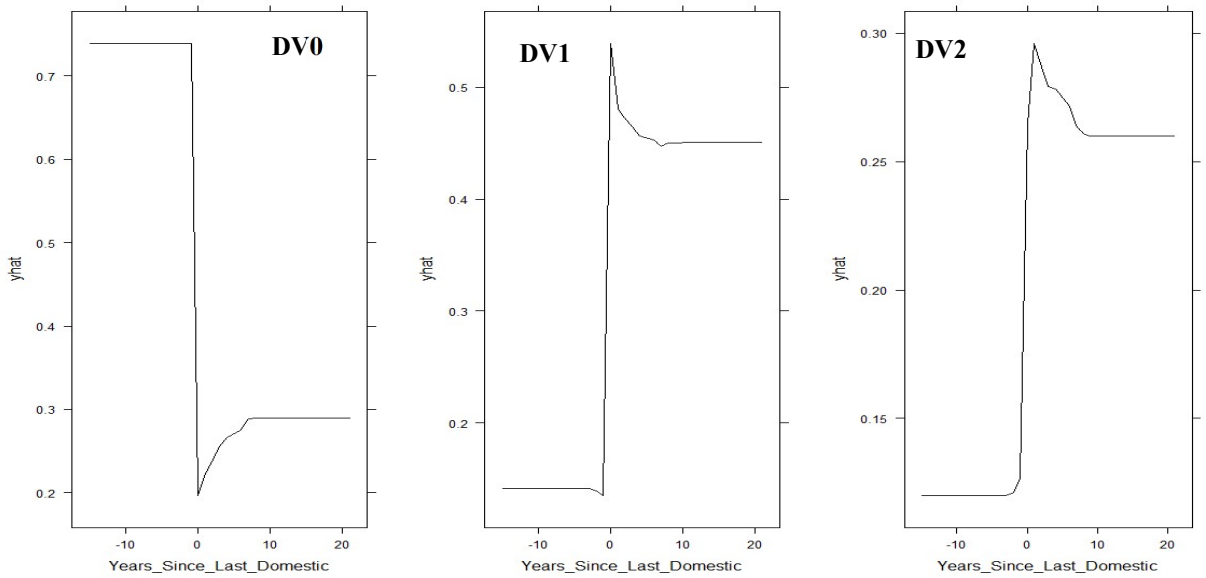

Figure 26. Partial response plots for years since last arrested for a domestic crime[24]

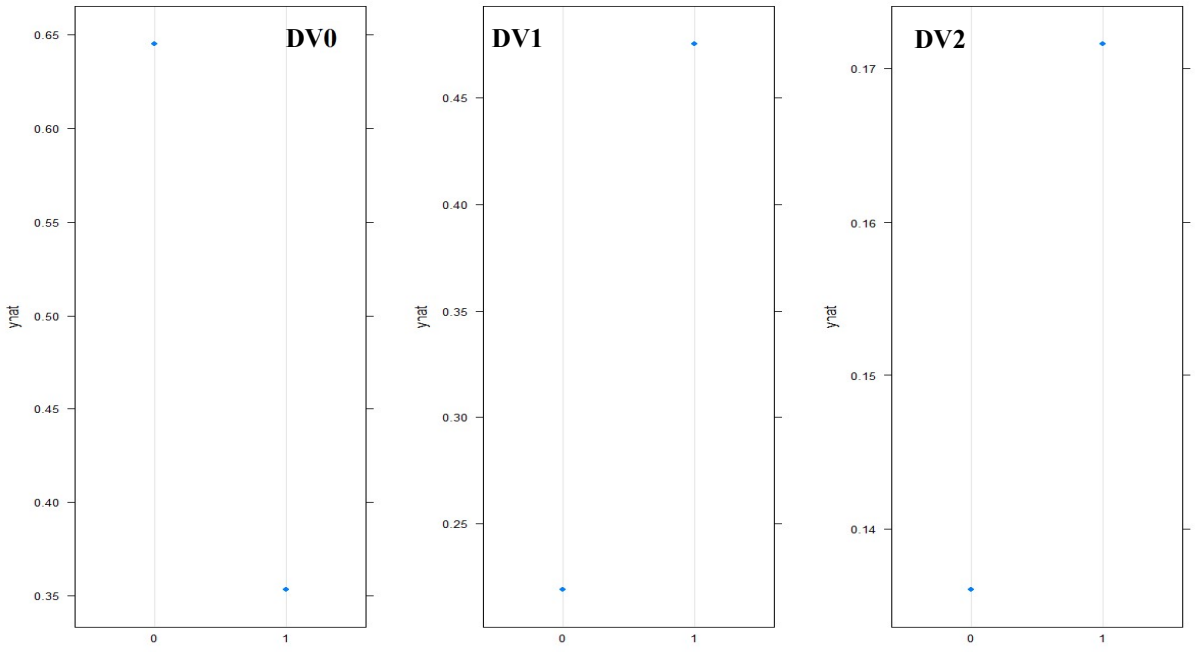

Figure 27. Partial response plots for presenting arrest was for a domestic crime

[24] Note that the scale goes to –10 on these plots due to one erroneous record which had a misclassified date of crime and was missed in the cleaning process.

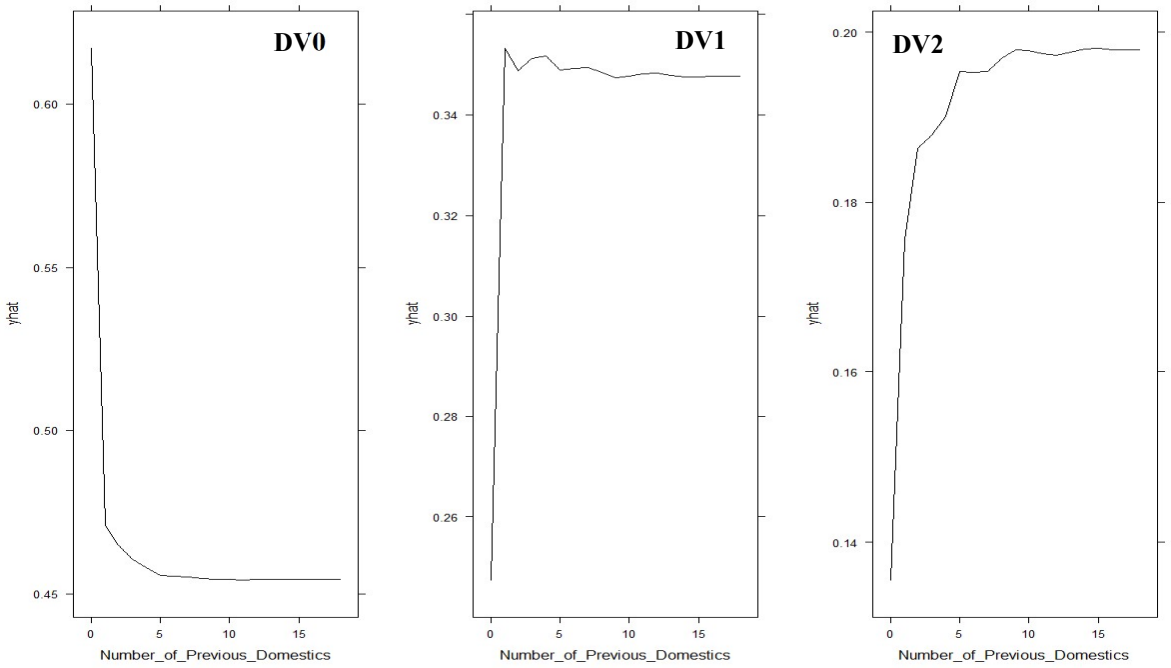

Figure 28. Partial response plots for number of previous domestic arrests

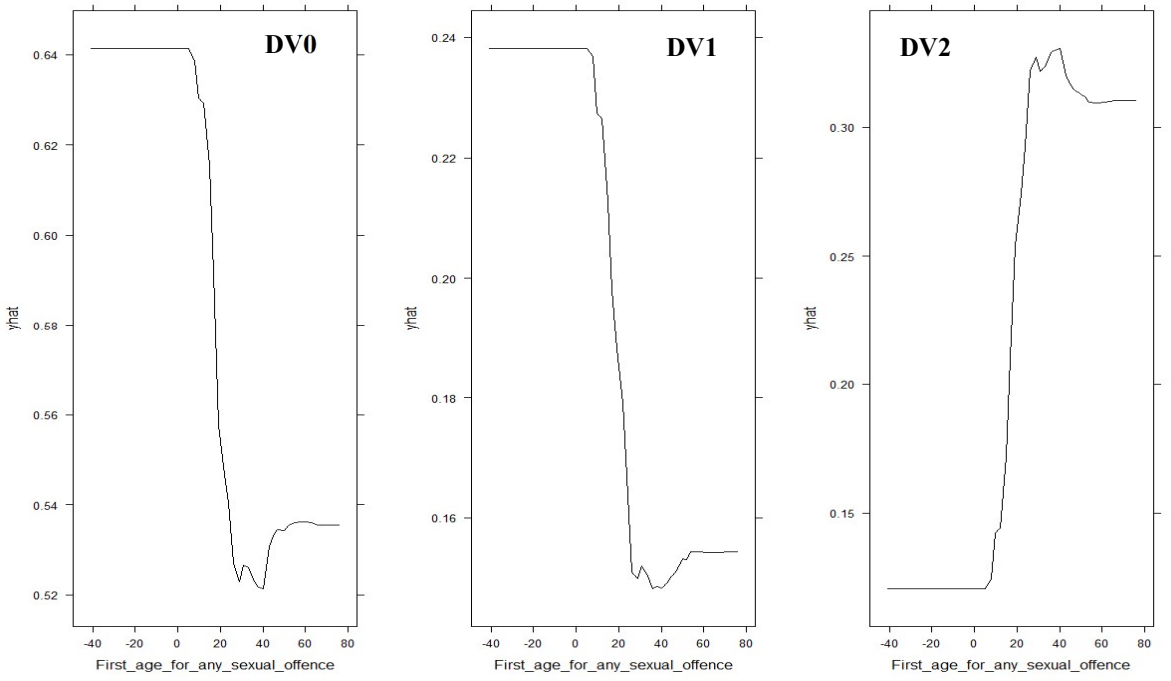

Figure 29. Partial response plots for age at first arrest for a sexual offence

23 Bibliography

Adler, M. J. (2001). *What is Crime?: Controversies Over the Nature of Crime and what to Do about it*. Rowman & Littlefield.

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G. and Rush, J.D., 2006. The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counselling Psychologist*, *34*(3), pp. 341-382.

Akman, D., & Normandeau, A., 1968. The measurement of crime and delinquency in Canada: A replication study. Acta Criminologica, 1, 135–260.

Aldarondo, E., & Mederos, F. (Eds.), 2002. *Programs for men who batter: Intervention and prevention strategies in a diverse society*. Civic Research Institute, Inc.

Ames, A., Di Antonio, E., Hitchcock, J., Webster, S., Wong, K., Ellingworth, D., Meadows, L., MacAlonan, D., Uhrig, N. and Logue, N., 2018. Adult Out of Court Disposal Pilot Evaluation-Final Report.

Andersen, H.A. and Mueller-Johnson, K., 2018. The Danish Crime Harm Index: How it works and why it matters. *Cambridge Journal of Evidence-Based Policing*, *2*(1-2), pp.52-69.

Anderson, M. A., Gillig, P. M., Sitaker, M., McCloskey, K., Malloy, K., Grigsby, N., 2003. "Why doesn't she just leave?" A descriptive study of victim reported impediments to her safety. *Journal of Family Violence*, 18, 151-155.

Ariel, B. and Bland, M., 2019. Is Crime Rising or Falling? A Comparison of Police Recorded Crime and Victimisation Surveys. *Methods of Criminology and Criminal Justice Research. (Sociology of Crime, Law, and Deviance)*, 24 pp.7-31.

Ariel, B., Weinborn, C. and Boyle, A., 2015. Can routinely collected ambulance data about assaults contribute to reduction in community violence?. *Emerg Med J*, *32*(4), pp.308-313.

Ashby, M.P., 2017. Comparing methods for measuring crime harm/severity. *Policing: A Journal of Policy and Practice*, *12*(4), pp.439-454.

Babyak, C., Alavi, A., Collins, K., Halladay, A., Tapper, D., 2009. *The Methodology of the Police-Reported Crime Severity Index*. Ottawa:, Statistics Canada(Catalogue no. HSMD-2009–006E/F).

Babyak, C., Campbell, A., Evra, R. and Franklin, S., 2013. *Updating the Police-reported Crime Severity Index Weights: Refinements to the Methodology*. Statistics Canada Catalogue no. HSMD-2013-005E/F.

Banerjee, A., Chitnis, U.B., Jadhav, S.L., Bhawalkar, J.S. and Chaudhury, S., 2009. Hypothesis testing, type I and type II errors. *Industrial psychiatry journal*, *18*(2), p.127.

Baraniuk, C., 2017. Durham Police AI to help with custody decisions. [online] bbc.co.uk. Available at https://www.bbc.co.uk/news/technology-39857645 [accessed 19 February 2019].

Barnes, G. and Hyatt, J.M., 2012. Classifying adult probationers by forecasting future offending.

Barnham, L., Barnes, G. C., & Sherman, L. W., 2017. Targeting escalation of intimate partner violence: evidence from 52,000 offenders. *Cambridge Journal of Evidence-Based Policing*, 1-27

Barnish, M., 2004. *Domestic violence: A literature review: summary*. HM Inspectorate of Probation.

Baker, N. V., Gregware, P. R., & Cassidy, M. A., 1999. Family killing fields: Honor rationales in the murder of women. *Violence against women*, 5(2), 164-184.

Barnard, G.W., Vera, H., Vera, M.I. and Newman, G., 1982. Till death do us part: A study of spouse murder. *Journal of the American Academy of Psychiatry and the Law Online*, *10*(4), pp.271-280.

Beck, U., Lash, S. and Wynne, B., 1992. *Risk society: Towards a new modernity* (Vol. 17). Sage.

Berk, R., 2012. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Berk, R.A. and Bleich, J., 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y*, *12*, p.513.

Berk, R.A., He, Y. and Sorenson, S.B., 2005. Developing a practical forecasting screener for domestic violence incidents. *Evaluation Review*, *29*(4), pp.358-383.

Berk, R.A., Kriegler, B. and Baek, J.H., 2006. Forecasting dangerous inmate misconduct: An application of ensemble statistical procedures. *Journal of Quantitative Criminology*, *22*(2), pp.131-145.

Berk, R., Sherman, L., Barnes, G., Kurtz, E. and Ahlman, L., 2009. Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), pp.191-211.

Berk, R.A., Sorenson, S.B. and Barnes, G., 2016. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, *13*(1), pp.94-115.

Berry, V., Stanley, N., Radford, L., McCarry, M. and Larkins, C., 2014. *Building effective responses: an independent review of violence against women, domestic abuse and sexual violence services in Wales*.

Best, J., & Luckenbill, D. F., 1996. Careers in deviance and respectability. In D. F. Greenberg (Ed.), *Criminal Careers* (pp. 3–14). Brookfield, VT: Dartmouth.

Bland, M., & Ariel, B., 2015. Targeting escalation in reported domestic abuse: Evidence from 36,000 callouts. *International Criminal Justice Review*, 25(1), 30–53. doi:10.1177/1057567715574382

Blumstein, A. 1974. Seriousness weights in an index of crime. *American Sociological Review*, 39, 854–864.

Blum-West, S. R. 1985. The seriousness of crime: A study of popular morality. *Deviant Behavior*, 6, 83–98.

Bocko, S., Cicchetti, C., Lempicki, L. and Powell, A., 2004. Restraining order violators, corrective programming and recidivism. *Boston, MA: Office of the Commissioner of Probation*.

Bond, E. and Tyrrell, K., 2018. Understanding revenge pornography: A national survey of police officers and staff in England and Wales. *Journal of interpersonal violence*, p.0886260518760011.

Bonta, J., 1996. Risk needs assessment and treatment. In A.T.Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp.18-32). Thousand Oaks, CA: Sage.

Bottoms, A., Shapland, J., Costello, A., Holmes, D., & Muir, G. (2004). Towards desistance: Theoretical underpinnings for an empirical study. *The Howard Journal of Crime and Justice,* 43(4), 368-389.

Bottoms, A. and Tankebe, J., 2012. Beyond procedural justice: A dialogic approach to legitimacy in criminal justice. *J. Crim. L. & Criminology*, *102*, p.119.

Box, S. 1983. *Power, Crime, and Mystification*, London: Tavistock.

Brand, S. and Price, R. 2000. *The economic and social costs of crime*, Home Office Research Study 217. London: Home Office Research, Development and Statistics Directorate.

Breiman L. 2001. Random forests. *Machine Learning* 45:5–32

Bridger, E., Strang, H., Parkinson, J. and Sherman, L.W., 2017. Intimate partner homicide in England and Wales 2011–2013: Pathways to prediction from multi-agency domestic homicide reviews. *Cambridge Journal of Evidence-Based Policing*, *1*(2-3), pp.93-104.

Bridges, G. S., & Lisagor, N. S., 1975. Scaling seriousness: An evaluation of magnitude and category scaling techniques. *Journal of Criminal Law and Criminology*, 66, 215–221.

Burgess, E. W., 1928. Factors determining success or failure on parole, Part IV of A.A. Bruce et al., *The Workings of the Indeterminate Sentence Law and the Parole System in Illinois*. Springfield, IL: The Board of Parole.

Brimicombe, A.J., 2018. Mining police-recorded offence and incident data to inform a definition of repeat domestic abuse victimization for statistical reporting. *Policing: A Journal of Policy and Practice*, *12*(2), pp.150-164.

Brimicombe, A.J., Brimicombe, L.C. and Li, Y., 2007. Improving geocoding rates in preparation for crime data analysis. *International Journal of Police Science & Management*, *9*(1), pp.80-92.

Brisson, N. J., 1981. Battering husbands: A survey of abusive men. Victimology: An International Journal, 6, 338–344.

Brooks-Hay, O. and Burman, M., 2018. *Domestic Abuse: Contemporary perspectives and innovative practices*. Dunedin Academic Press Ltd.

Burgess, M., 2018. *UK police are using AI to inform custodial decisions – but it could be discriminating against the poor*. [ONLINE] Available at: https://www.wired.co.uk/article/police-ai-uk-durham-hart-checkpoint-algorithm-edit. [Accessed 15 January 2019]

Buzawa, E., Hotaling, G., Klein, A. and Byrnes, J., 1999. Response to Domestic Violence in a Pro-Active Court Setting, Final Report. Washington DC: US Department of Justice.

Camacho, C.M. and Alarid, L.F., 2008. The significance of the victim advocate for domestic violence victims in municipal court. *Violence and Victims*, *23*(3), pp.288-300.

Campbell, J. C., 1995. *Assessing dangerousness: Violence by sexual offenders, batterers, and child abusers.* Newbury Park, CA: Sage

Campbell, J.C., 2005. Assessing dangerousness in domestic violence cases: History, challenges, and opportunities. *Criminology & Public Policy*, *4*(4), pp.653-672.

Campbell, J.C., 2007. Prediction of homicide of and by battered women. *Assessing dangerousness: Violence by batterers and child abusers*, *2*.

Campbell, J. C., Glass, N., Sharps, P. W., Laughon, K., & Bloom, T. (2007). Intimate partner homicide: review and implications of research and policy. *Trauma, Violence, & Abuse*, 8(3), 246-269.

Campbell, J.C., Sharps, P. and Glass, N., 2001. Risk assessment for intimate partner homicide.

Cantos, A.L., Goldstein, D.A., Brenner, L., O'Leary, K.D. and Verborg, R., 2015. Correlates and Program Completion of Family Only and Generally Violent Perpetrators of Intimate Partner Violence. *Behavioural Psychology/Psicologia Conductual*, *23*(3).

Cantos, A.L. and O'Leary, K.D., 2014. One size does not fit all in treatment of intimate partner violence. *Partner Abuse*, *5*(2), pp.204-236.

Carrell, S.E. and Hoekstra, M., 2012. Family business or social problem? The cost of unreported domestic violence. *Journal of Policy Analysis and Management*, *31*(4), pp.861-875.

Carlo, S., *Artificial Intelligence, Big Data and the Rule of Law,* Event Report, The Bingham Centre for the Rule of Law, 9 October 2017 https://www.biicl.org/event/1280.

Cavanaugh, M.M. and Gelles, R.J., 2005. The utility of male domestic violence offender typologies: New directions for research, policy, and practice. *Journal of interpersonal violence*, *20*(2), pp.155-166.

Chalkley, R. and Strang, H., 2017. Predicting domestic homicides and serious violence in Dorset: A replication of Thornton's Thames Valley analysis. *Cambridge Journal of Evidence-Based Policing*, *1*(2-3), pp.81-92.

Chambers-McClellan, A., 2002. *Evidence for the escalation of domestic violence in 911 call records* (Doctoral dissertation, Medical College of Georgia).

Chen, Y.S., Chong, P.P. and Tong, Y., 1993. Theoretical foundation of the 80/20 rule. *Scientometrics*, *28*(2), pp.183-204.

Cho, H. and Wilke, D.J., 2010. Gender differences in the nature of the intimate partner violence and effects of perpetrator arrest on revictimization. *Journal of family violence*, *25*(4), pp.393-400.

Cockbain, E. and Knutsson, J. eds., 2014. *Applied police research: Challenges and opportunities*. Routledge.

Cohen, H. and Mandrack, M.M., 2002. Application of the 80/20 rule in safeguarding the use of high-alert medications. *Critical care nursing clinics of North America*, *14*(4), pp.369-374.

Cohen, M. A., 1988. Some new evidence on the seriousness of crime. *Criminology*, 26, 343–353.

College of Policing., 2018. *Domestic abuse index***.** [ONLINE] Available at: https://www.app.college.police.uk/domestic-abuse-index/. [Accessed 01 March 2019]

Coughlan, S. 2019. *School spending on pupils cut by 8%, says IFS*. [ONLINE] Available at: https://www.bbc.co.uk/news/education-44794205. [Accessed 7 January 2019].

Crawford, M., & Gartner, R. (1992). *Women killing: Intimate femicide in Ontario, 1974-1990*. Toronto, Ontario, Canada: Women's Directorate, Ministry of Social Services.

Curtis-Ham, S. and Walton, D., 2017. The New Zealand crime harm index: Quantifying harm using sentencing data. *Policing: A Journal of Policy and Practice*, *12*(4), pp.455-467.

Davies, P. A. and Biddle, P., 2018. 'Implementing a perpetrator-focused partnership approach to tackling domestic abuse: The opportunities and challenges of criminal justice localism', *Criminology & Criminal Justice*, 18(4), pp. 468–487.

Dawes, R.M., Faust, D. and Meehl, P.E., 1989. Clinical versus actuarial judgment. *Science*, *243*(4899), pp.1668-1674.

Dawson, M. and Dinovitzer, R., 2001. Victim cooperation and the prosecution of domestic violence in a specialized court. *Justice Quarterly*, *18*(3), pp.593-622.

Dixon, L. and Browne, K., 2003. The heterogeneity of spouse abuse: A review. *Aggression and violent behavior*, *8*(1), pp.107-130.

Dodd, V., 2018. *England and Wales police funding rise of £970m 'not enough'.* [ONLINE] Available at: https://www.theguardian.com/uk-news/2018/dec/13/england-and-wales-police-funding-rise-of-970m-not-enough [Accessed 14 January 2019].

Dolan, M. and Doyle, M., 2000. Violence risk prediction: Clinical and actuarial measures and the role of the Psychopathy Checklist. *The British Journal of Psychiatry*, *177*(4), pp.303-311.

Dolan, P. and Peasgood, T., 2007. Estimating the economic and social costs of the fear of crime, *British Journal of Criminology*, vol. 47(1), pp 121-132. Oxford University Press, Oxford.

Dubourg, R., Hamed, J. and Thorns, J., 2005. *The economic and social costs of crime against individuals and households 2003/04*, Home Office Online Report, 30/05.

Dudfield, G., Angel, C., Sherman, L.W. and Torrence, S., 2017. The "power curve" of victim harm: Targeting the distribution of crime harm index values across all victims and repeat victims over 1 year. *Cambridge Journal of Evidence-Based Policing*, *1*(1), pp.38-58.

Dutton, D. G., and Kerry, G. (2002). Modus operandi and personality disorders in incarcerated spousal killers. *Journal of Psychiatric Practice*, 8(4), 216-228.

Dutton, D.G. and Kropp, P.R., 2000. A review of domestic violence risk instruments. *Trauma, violence, & abuse*, *1*(2), pp.171-181.

Eck, J. E., 2003. Preventing crime at places. In *Evidence-based crime prevention* (pp. 255-308). Routledge.

Edelstein, A., 2016. Rethinking conceptual definitions of the criminal career and serial criminality. *Trauma, Violence, & Abuse*, *17*(1), pp.62-71.

Egger, Steven A. 1985. *"An Analysis of the Serial Murder Phenomenon and the Law Enforcement Response,"*, Ph.D. dissertation Sam Houston State University.

Elbow, M., 1977. Theoretical considerations of violent marriages. *Social casework*, *58*(9), pp.515-526.

Elisha, E., Idisis, Y., Timor, U., & Addad, M. (2010). Typology of intimate partner homicide: Personal, interpersonal, and environmental characteristics of men who murdered their female intimate partner. *International journal of offender therapy and comparative criminology*, 54(4), 494-516.

Epperlein, T. and Nienstedt, B.C., 1989. Reexamining the use of seriousness weights in an index of crime. *Journal of Criminal justice*, *17*(5), pp.343-360.

Evans, M., 2016. The most severe crimes up by 30 per cent in some areas. [ONLINE] Available at https://www.telegraph.co.uk/news/2016/11/30/severe-crimes-30-per-cent-areas/ [accessed 4th March 2019].

Evans, S.S. and Scott, J.E., 1984. Effects of item order on the perceived seriousness of crime: A reexamination. *Journal of research in crime and delinquency*, *21*(2), pp.139-151.

Feld, S. L., & Straus, M. A. (1989). Escalation and desistance of wife assault in marriage. *Criminology*, 27(1), 141-162.

Feld, S.L., & Straus, M. A. (1990). 'Escalation and desistance of wife assault in marriage' in M. A Straus and R. J. Gelles edited with the assistance of C. Smith, *Physical violence in American families: Risk factors and adaptations to violence in 8,145 families* (pp. 489505). New Brunswick, NJ: Transaction Publishers.

Felson, R., Ackerman, J. and Gallagher, C., 2005. Police intervention and the repeat of domestic assault. *Criminology* 43(3): 563-588.

Felson, R. B., & Paré, P. P. (2005). The reporting of domestic violence and sexual assault by nonstrangers to the police. *Journal of marriage and family*, 67(3), 597-610.

Figlio, R. M. (1975). The seriousness of offenses: An evaluation by offenders and nonoffenders. *Journal of Criminal Law and Criminology*, 66, 189–200.

Fishman, G., Kraus, V. and Cohen, B.Z., 1986. A multidimensional approach to the problem of crime seriousness. *International Journal of Comparative and Applied Criminal Justice*, *10*(1-2), pp.177-191.

Fitz-Gibbon, K. and Walklate, S., 2017. The efficacy of Clare's Law in domestic violence law reform in England and Wales. *Criminology & Criminal Justice*, *17*(3), pp.284-300.

Fixsen, D.L., Blase, K.A., Naoom, S.F. and Wallace, F., 2009. Core implementation components. *Research on social work practice*, *19*(5), pp.531-540.

Fleming, S. (1981). The closed mind and the judgement of crime: A replication of the Sellin-Wolfgang index. *International Journal of Comparative and Applied Criminal Justice*, 5, 51–64.

Florence, C., Shepherd, J., Brennan, I. and Simon, T., 2011. Effectiveness of anonymised information sharing and use in health service, police, and local government partnership for preventing violence related injury: experimental study and time series analysis. *British Medical Journal*, *342*, p.d3313.

Frieze, I. H., & Browne, A. 1989. Violence in marriage. In L. E. Ohlin & M. H. Tonry (Eds.), *Family violence*. Chicago: University of Chicago Press.

Giles-Sims, J., 1983. *Wife battering: a systems theory approach*. New York: Guilford Press.

Goldstein, H., 1990. Excellence in problem-oriented policing. *New York NY*.

Gondolf, E. W. (1988). Who are these guys? Toward a behavioural typology of batterers. *Violence and Victims*, 3, 187–203.

Gondolf, E.W., 1998. The victims of court-ordered batterers: Their victimization, helpseeking, and perceptions. *Violence Against Women*, *4*(6), pp.659-676.

Goosey, J., Sherman, L. and Neyroud, P., 2017. Integrated case management of repeated intimate partner violence: a randomized, controlled trial. *Cambridge Journal of Evidence-Based Policing*, *1*(2-3), pp.174-189.

Gottfredson, S.D. and Moriarty, L.J., 2006. Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, *52*(1), pp.178-200.

Gottfredson, S. D., Young, K. L., & Laufer, W. S., 1980. Additivity and interactions in offense seriousness scales. *Journal of Research in Crime and Delinquency*, 17, 26–41.

Gottman, J. M., Jacobson, N. S., Rushe, R. H., Shortt, J. W., Babcock, J., La Taillade, J. J., & Waltz, J., 1995. The relationship between heart rate reactivity, emotionally aggressive behaviour and general violence in batterers. *Journal of Family Psychology*, 9(3), 227–248

Goussinsky, R., & Yassour-Borochowitz, D., 2012. "I killed her, but I never laid a finger on her"—A phenomenological difference between wife-killing and wife-battering. *Aggression and Violent Behavior*, 17(6), 553-564.

Gracia, E. (2004). Unreported cases of domestic violence against women: towards an epidemiology of social silence, tolerance, and inhibition. *Journal of Epidemiology and Community Health*, 2004, 58 (7): 536-537.

Greenfield, V.A. and Paoli, L., 2013. A framework to assess the harms of crimes. *British Journal of Criminology*, *53*(5), pp.864-885.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. and Nelson, C., 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1), p.19.

Hamberger, L.K. and Hastings, J.E., 1988. Skills training for treatment of spouse abusers: An outcome study. *Journal of Family Violence*, *3*(2), pp.121-130.

Hamberger, L. K., Lohr, J. M., Bonge, D., & Tolin, D. F., 1996. A large sample empirical typology of male spouse abusers and its relationship to dimensions of abuse. Violence and Victims, 11, 277–292.

Hanna, C., 1996. No right to choose: Mandated victim participation in domestic violence prosecutions. *Harvard law review*, pp.1849-1910.

Hansel, M., 1987. Citizen crime stereotypes—Normative consensus revisited. *Criminology*, 25, 455–485.

Harcourt, B.E., 2014. Risk as a proxy for race: The dangers of risk assessment. *Fed. Sent'g Rep.*, *27*, p.237.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning: prediction, inference and data mining. *Springer-Verlag, New York*.

Hazelwood, R., & Burgess, A. W., 1987. An introduction to the serial rapist: Research by the FBI. *FBI Law Enforcement Bulletin*, 16-24.

Heeks, M., Reed, S., Tafsiri, M. and Prince, S., 2018. The economic and social costs of crime Second edition.

Henry, N. and Powell, A. eds., 2014. *Preventing sexual violence: Interdisciplinary approaches to overcoming a rape culture*. Springer.

Her Majesty's Inspectorate of the Constabulary, Fire and Rescue Services, 2014a. *Everyone's business: Improving the police response to domestic violence*. [Online] Retrieved from https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/2014/04/improving-the-police-response-to-domestic-abuse.pdf [accessed 15th October 2016]

Her Majesty's Inspectorate of the Constabulary, Fire and Rescue Services, 2014b. *Crime recording: making the victim count*. [Online] Retrieved from https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/crime-recording-making-the-victim-count.pdf. [accessed 15th October 2016]

Her Majesty's Inspectorate of the Constabulary, Fire and Rescue Services., 2015. *Increasingly everyone's business: A progress report on the police response to domestic abuse*. [online] https://www.justiceinspectorates.gov.uk/hmicfrs/publications/increasingly-everyones-business-a-progress-report-on-the-police-response-to-domestic-abuse/ [accessed 7th May 2017]

Her Majesty's Inspectorate of Constabulary, Fire and Rescue Services., 2017. *A progress report on the police response to domestic abuse.* [Online]. https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/progress-report-on-the-police-response-to-domestic-abuse.pdf [accessed 9th February, 2019].

Her Majesty's Inspectorate of Constabulary, Fire and Rescue Services., 2019. *A progress report on the police response to domestic abuse* [Online] https://www.justiceinspectorates.gov.uk/hmicfrs/publications/a-progress-report-on-the-police-response-to-domestic-abuse/ [accessed 4th March 2019].

Hester, M., 2013. Who does what to whom? Gender and domestic violence perpetrators in English police records. *European Journal of Criminology*, 10(5), 623-637.

Hester, M. and Westmarland, N., 2006. Domestic violence perpetrators. *Criminal Justice Matters*, *66*(1), pp.34-35.

Hillyard, P. and Tombs, S., 2007. From 'crime'to social harm?. *Crime, law and social change*, *48*(1-2), pp.9-25.

Holmes, R.M. and DeBurger, J.E., 1998. Profiles in terror: The serial murderer. *Contemporary perspectives on serial murder*, pp.5-16.

Holtzworth-Munroe, A. and Stuart, G.L., 1994. Typologies of male batterers: Three subtypes and the differences among them. *Psychological bulletin*, *116*(3), p.476.

Home Affairs Select Committee, 2009. Young black people and the criminal justice system [Online]. https://dera.ioe.ac.uk/1043/1/young-black-people-cjs-dec09.pdf [accessed 12th January 2019]

Home Office (2012, 18 September). *New definition of domestic violence*. Retrieved from https://www.gov.uk/government/news/new-definition-of-domestic-violence

Horning, N., 2013. Introduction to decision trees and random forests. *Am. Mus. Nat. Hist*, *2*, pp.1-27.

Hotton, T., 2001. *Spousal violence after marital separation*. Canadian Centre for Justice Statistics.

Hoyle, C., 2008. Will she be safe? A critical analysis of risk assessment in domestic violence cases. *Children and Youth Services Review*, *30*(3), pp.323-337.

House, P.D. and Neyroud, P.W., 2018. Developing a Crime Harm Index for Western Australia: the WACHI. *Cambridge Journal of Evidence-Based Policing*, *2*(1-2), pp.70-94.

Hulme, W., 2017. Local authorities' budgets are roughly 26% lower since 2010. [Online] https://fullfact.org/economy/local-authorities-budgets/ [accessed October 26th 2018]

Ignatans, D. and Pease, K., 2015. Taking crime seriously: playing the weighting game. *Policing: a Journal of Policy and Practice*, *10*(3), pp.184-193.

Iqbal, M. and Rizwan, M., 2009, August. Application of 80/20 rule in software engineering Waterfall Model. In *2009 International Conference on Information and Communication Technologies* (pp. 223-228). IEEE.

Iyengar, R., 2009. Does the certainty of arrest reduce domestic violence? Evidence from mandatory and recommended arrest laws. *Journal of public Economics*, *93*(1-2), pp.85-98.

Jacobson, N.S. and Gottman, J.M., 1998. *When men batter women: New insights into ending abusive relationships*. Simon and Schuster.

Johnson, M. P. 1995. Patriarchal terrorism and common couple violence: Two forms of violence against women. *Journal of Marriage and the Family*, 57, 283–294.

Johnson, M.P. and Ferraro, K.J., 2000. Research on domestic violence in the 1990s: Making distinctions. *Journal of Marriage and Family*, *62*(4), pp.948-963.

Johnson, M. P., 2006. Conflict and control: Gender symmetry and asymmetry in domestic violence. *Violence Against Women*, 12, 1003-1018.

Johnson, C., & Sachmann, M. 2014. Familicide-Suicide: From Myth To Hypothesis And Toward Understanding. *Family Court Review*, 52(1), 100-113.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), pp.255-260.

Kahneman, D. and Klein, G., 2009. Conditions for intuitive expertise: a failure to disagree. *American psychologist*, *64*(6), p.515.

Kasturirangan, A., Krishnan, S., & Riger, S., 2004. The impact of culture and minority status on women's experience of domestic violence. *Trauma, Violence, & Abuse*, 5(4), 318-332.

Kaye, M., Stubbs, J. and Tolmie, J., 2003. Domestic violence, separation and parenting: Negotiating safety using legal processes. *Current Issues in Criminal Justice*, *15*(2), pp.73-94.

Kelly, L., Adler, J.R., Horvath, M.A., Lovett, J., Coulson, M., Kernohan, D. and Gray, M., 2013. *Evaluation of the Pilot of Domestic Violence Protection Orders*. Home Office Science Research Paper 76: (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/260897/horr76. pdf)

Kerr, J., Whyte, C., & Strang, H., 2017. Targeting Escalation and Harm in Intimate Partner Violence: Evidence from Northern Territory Police, Australia. *Cambridge Journal of Evidence-Based Policing*, 1-17.

Klein, A., 1996. Reabuse in a population of court restrained male batterers. In E. Buzawa & C. Buzawa (Eds.), *Do arrest and restraining orders work*? (pp. 192-214). Thousand Oaks, CA: Sage.

Klein, A., & Tobin, T., 2008. Longitudinal study of arrested batterers, 1995–2005: Career criminals. *Violence Against Women*, 14(2), 136–157.

Klein, A., Wilson, D., Crowe, A., & DeMichele, M., 2005. *Evaluation of the Rhode Island probation specialized domestic violence supervision unit* [NCJ 222912]. Retrieved from http://www.ncjrs.gov/ App/Publications/abstract.aspx?ID=244821

Klevens, J., Baker, C.K., Shelley, G.A. and Ingram, E.M., 2008. Exploring the links between components of coordinated community responses and their impact on contact with intimate partner violence services. *Violence against women*, *14*(3), pp.346-358.

Klevens, J. and Cox, P., 2008. Coordinated community responses to intimate partner violence: Where do we go from here. *Criminology & Pub. Pol'y*, *7*, p.547.

Kocsis, R.N., Cooksey, R.W. and Irwin, H.J., 2002. Psychological profiling of offender characteristics from crime behaviors in serial rape offences. *International Journal of Offender Therapy and Comparative Criminology*, *46*(2), pp.144-169.

Kocsis, R.N. and Irwin, H.J., 1998. The psychological profile of serial offenders and a redefinition of the misnomer of serial crime. *Psychiatry, Psychology and Law*, *5*(2), pp.197-213.

Kock, R., 1999. *80-20 principle: The secret to success by achieving more with less*. New York: Doubleday.

Kraemer, F., Van Overveld, K. and Peterson, M., 2011. Is there an ethics of algorithms?. *Ethics and Information Technology*, *13*(3), pp.251-260.

Kropp, P.R., 2004. Some questions regarding spousal assault risk assessment. *Violence against women*, *10*(6), pp.676-697.

Kulkarni, V.Y. and Sinha, P.K., 2012, July. Pruning of random forest classifiers: A survey and future directions. In *2012 International Conference on Data Science & Engineering (ICDSE)* (pp. 64-68). IEEE.

Kwan, L., 2016. *Western Australian maximum sentence values compared with the Cambridge Crime Harm Index*, Internship report.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.

Liberty., 2019. *Liberty report exposes police forces' use of discriminatory data to predict crime*. [Online]. https://www.libertyhumanrights.org.uk/news/press-releases-and-statements/liberty-report-exposes-police-forces'-use-discriminatory-data-0 [accessed 4th March 2019].

Litwack, T.R., 2001. Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, *7*(2), p.409.

Litwack, T. R., & Schlesinger, L. B. (1999). Dangerousness risk assessments: Research, legal, and clinical considerations. In A. Hess & I. Weiner (Eds.), *The handbook of forensic psychology* (pp. 171–217). New York: Wiley.

Liu, Y.Y., Yang, M., Ramsay, M., Li, X.S. and Coid, J.W., 2011. A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*, *27*(4), pp.547-573.

Lloyd, S., Farrell, G. and Pease, K., 1994. *Preventing repeated domestic violence: A demonstration project on Merseyside*. London: Home Office Police Research Group.

Loinaz, I., 2014. Typologies, risk and recidivism in partner-violent men with the B-SAFER: A pilot study. *Psychology, Crime & Law*, *20*(2), pp.183-198.

Lum, C. M., & Koper, C. S., 2017. *Evidence-based policing: Translating research into practice*. Oxford: Oxford University Press.

Lynch, J. P., & Danner, M. J. E., 1993. Offense seriousness scaling: An alternative to scenario methods. *Journal of Quantitative Criminology*, 9, 309–322.

Malach-Pines, A., 2002. *Falling in love: How we choose the lovers we choose*. New York: Taylor & Francis Group.

Malmquist, C. P., 2007. *Homicide: A psychiatric perspective*. American Psychiatric Pub.

Maxwell, Christopher D., Joel H. Garner, and Jeffrey A. Fagan. "The preventive effects of arrest on intimate partner violence: Research, policy and theory." *Criminology & Public Policy* 2, no. 1 (2002): 51-80.

McCleary, R., O'Neil, M. J., Epperlein, T., Jones, C., & Gray, R. H., 1981. Effects of legal education and work experience on perceptions of crime seriousness. *Social Problems*, 28, 276–289.

McLaughlin, H., Banks, C., Bellamy, C., Robbins, R. and Thackray, D., 2014. *Domestic violence, adult social care and MARACs: implications for practice*, NHS National Institute for Health Research, Research Findings [Online] accessed 28th November 2017. Available from: http://www.sscr.nihr.ac.uk/PDF/Findings/RF44.pdf

Meehl, P., 1954. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Miethe, T. D., 1991. Social psychophysical measurement: A comparison of the measurement properties of magnitude and categorical scaling of social perceptions. *Social Science Quarterly*, 67, 195–204.

Miller, T.R., Cohen, M.A. and Wiersema, B., 1996. *Victim costs and consequences: A new look*. Washington, DC: US Department of Justice, Office of Justice Programs, National Institute of Justice.

Mills, L.G., Barocas, B. and Ariel, B., 2013. The next generation of court-mandated domestic violence treatment: A comparison study of batterer intervention and restorative justice programs. *Journal of Experimental Criminology*, *9*(1), pp.65-90.

Mintz, E. (1980). Obsession with the rejecting beloved. *Psychoanalytic Review*, 67, 479-492

Mitchell, B. A. (1997). *The etiology of serial murder: Towards an integrated model*. Cambridge, England: University of Cambridge.

Mitchell, R. J. (2016). The Sacramento hot spots policing experiment: An extension and sensitivity analysis. *Unpublished dissertation, University of Cambridge, Cambridge*.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L., 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), p.2053951716679679.

Moffitt, T. E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological review*, 100(4), 674.

Morgan, R., Maguire, M. and Reiner, R. eds., 2012. *The Oxford handbook of criminology*. Oxford University Press.

Myhill, A., 2015. Measuring coercive control: What can we learn from national population surveys? *Violence against women*, 21(3), 355-375.

Myhill, A., 2018. *The police response to domestic violence: Risk, discretion, and the context of coercive control* (Doctoral dissertation, City, Universtiy of London).

Neyroud, P., 2012. Policing and ethics. In *Handbook of policing* (pp. 694-720). Willan.

Neyroud, P., 2015. Future Perspectives in Policing: A Crisis or a Perfect Storm: The Trouble with Public Policing? In *Police Services* (pp. 161-165). Springer, Cham.

Neyroud, P., 2017. *Learning to Field Test in Policing: Using an analysis of completed randomised controlled trials involving the police to develop a grounded theory on the factors contributing to high levels of treatment integrity in Police Field Experiments* (Doctoral dissertation, University of Cambridge).

Neyroud, P. and Disley, E., 2008. Technology and policing: Implications for fairness and legitimacy. *Policing: A Journal of Policy and Practice*, *2*(2), pp.226-232.

Neyroud, P.W. and Weisburd D., 2014. Transforming the police through science: the challenge of ownership. *Policing: A Journal of Policy and Practice, 287-293.*

Office for National Statistics (ONS), 2016a. *Domestic abuse in England and Wales: year ending March 2016.* Statistical Bulletin. London, UK: Office of National Statistics. [Online] Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwales/yearendingmarch2016 [accessed 17th March 2018]

Office for National Statistics (ONS), 2016b *Research outputs: developing a Crime Severity Score for England and Wales using data on crimes recorded by the police* [online] Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/researchoutputsdevelopingacrimeseverityscoreforenglandandwalesusingdataoncrimesrecordedbythepolice/2016-11-29 [accessed 6th March 2019]

Office for National Statistics (ONS). (2017). *Domestic abuse in England and Wales: year ending March 2017.* Statistical Bulletin. London, UK: Office of National Statistics. [Online] Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwales/yearendingmarch2017 [accessed 17th March 2018]

Office for National Statistics (ONS). (2018). *Domestic abuse in England and Wales: year ending March 2018.* Statistical Bulletin. London, UK: Office of National Statistics. [Online] Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwales/yearendingmarch2018 [accessed 2nd March 2019]

Office for National Statistics (ONS). (2019). *Police workforce, England and Wales: 30 September 2018.* Statistical Bullterin. London, UK: Office for National Statistics. [Online] Retrieved from https://www.gov.uk/government/statistics/police-workforce-england-and-wales-30-september-2018 [accessed 29th May 2019]

Okun, L., 1986. *Woman abuse: Facts replacing myths*. New York: Albany State University of New York Press.

Oswald, M., Grace, J., Urwin, S. and Barnes, G.C., 2018. Algorithmic risk assessment policing models: lessons from the Durham HART model and 'experimental' proportionality. *Information & Communications Technology Law*, *27*(2), pp.223-250.

Owens, Catherine, David Mann, and Roy Mckenna., (2014). *The Essex BWV Trial: The Impact of BWV on Criminal Justice Outcomes of Domestic Abuse Incidents*. London, UK: College of Policing

Pagelow, M. D. (1981). *Woman-battering: Victims and their experiences*. Beverly Hills, CA: Sage.

Paladin, (2014). *Serial perpetrator register and order.* [Online] https://paladinservice.co.uk/serial-perpetrator-register-and-order/ [accessed 4th March 2019]

Palle, C. and Godefroy, T., 2000. The cost of crime: A monetary assessment of offending in 1996. *Research on Crime and Criminal Justice in France: Penal Issues*.

Paoli, L., Greenfield, V. A., & Zoutendijk, A. (2013). The harms of cocaine trafficking: Applying a new framework for assessment. *Journal of Drug Issues*, *43*(4), 407-436.

Parton, D. A., Hansel, M., & Stratton, J. R. (1991). Measuring crime seriousness: Lessons from the National Survey of Crime Severity. *British Journal of Criminology*, 31, 72–85.

Pease, K., Ireson, J. and Thorpe, J., 1974. Additivity assumptions in the measurements of delinquency. *Brit. J. Criminology*, *14*, p.256.

Pennell, J. and Burford, G., 2000. Family group decision making: protecting children and women. *Child welfare*, *79*(2).

Petersson, J. and Strand, S., 2017. Recidivism in intimate partner violence among antisocial and family-only perpetrators. *Criminal Justice and Behaviour*, *44*(11), pp.1477-1495.

Piquero, A. R., Brame, R., Fagan, J., & Moffitt, T. E. (2006). Assessing the offending activity of criminal domestic violence suspects: Offense specialization, escalation, and de-escalation evidence from the Spouse Assault Replication Program. *Public Health Reports*, 121, 409.

Pontell, H. N., Granite, D., Keenan, C., & Geis, G., 1985. Seriousness of crimes: A survey of the nation's chiefs of police. *Journal of Criminal Justice*, 13, 1–13.

Polk, K., & Ranson, D. (1991). Role of gender in intimate homicide. *Australian and New Zealand Journal of Criminology*, 24(1), 15-24

Post, L.A., Klevens, J., Maxwell, C.D., Shelley, G. and Ingram, I., 2008 'The Impact of coordinated community response on communities' attitudes and rates of intimate partner violence', *American Journal of Public Health.*

Ptacek, J., 2017. Research on Restorative Justice in Cases of Intimate Partner Violence. *Preventing intimate partner violence: Interdisciplinary perspectives*, p.159.

Public Administration Select Committee., 2014. *Caught red-handed: why we can't count on police recorded crime statistics* [online]. https://publications.parliament.uk/pa/cm201314/cmselect/cmpubadm/760/760.pdf [accessed 4th March 2019.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk.* Washington, DC: American Psychological Association.

Ratcliffe, J.H., 2015. Towards an index for harm-focused policing. *Policing: A Journal of Policy and Practice*, *9*(2), pp.164-182.

Richards, L., 2006. Homicide Prevention: Findings from the Multi-agency Domestic Violence Homicide Review. *The Journal of Homicide and Major Incident Investigation*. Vol. 2 Issue 2. Autumn 2006. ACPO: Centrex

Richards, L., Letchford, S. and Stratton, S., 2008. *Policing Domestic Violence*. Oxford. Blackstone's Practical Policing, Oxford University Press.

Ridgeway, G., 2013. Linking prediction and prevention. *Criminology & Pub. Pol'y*, *12*, p.545.

Riedel, M. (1975). Perceived circumstances, inferences of intent, and judgments of offense seriousness. Journal of Criminal Law and Criminology, 66, 201–208.

Rinaldo, M.-B. V., 2015. *Comparing crime hotspots and crime harm-spots in a Swedish City: a descriptive analysis*. England: Cambridge University.

Rivas, C., Ramsay, J., Sadowski, L., Davidson, L.L., Dunne, D., Eldridge, S., Hegarty, K., Taft, A. and Feder, G., 2015. Advocacy interventions to reduce or eliminate violence and promote the physical and psychosocial well-being of women who experience intimate partner abuse. *Cochrane database of systematic reviews*, (12).

Robinson, A.L, 2016. What works for reducing domestic abuse: Risk-led policing and the DASH risk assessment tool [online] https://www.researchgate.net/profile/Amanda_Robinson5/publication/301821428_What_works_for_reducing_domestic_abuse_Risk-led_policing_and_the_DASH_risk_assessment_tool/links/5729c64208aef5d48d2ef55a/What-works-for-reducing-domestic-abuse-Risk-led-policing-and-the-DASH-risk-assessment-tool.pdf [accessed 3rd May 2017]

Robinson, A.L., 2017. Serial domestic abuse in Wales: an exploratory study into its definition, prevalence, correlates, and management. *Victims & Offenders*, *12*(5), pp.643-662.

Robinson, A.L., Myhill, A., Wire, J., Roberts, J. and Tilley, N., 2016. Risk-led policing of domestic abuse and the DASH risk model. *What Works: Crime Reduction Research. Cardiff & London: Cardiff University, College of Policing and UCL Department of Security and Crime Science*.

Rose, G. N. G., (1966), 'Concerning the Measurement of Delinquency', *British Journal of Criminology*, 6: 414—21.

Robinson, A.L. and Tregidga, J., 2005. *Domestic Violence MARACs (Multi-Agency Risk Assesment Conferences) for Very High-risk Victims in Cardiff, Wales: Views from the Victims*. Cardiff University School of Social Sciences.

Rossi, P. H., and Henry, J. P. (1980). Seriousness: A measure for all purposes? In M. W. Klein & K. S. Teilmann (Eds.), *Handbook of criminal justice evaluation* (pp. 489–505). Newbury Park: Sage.

Rossi, P.H., Simpson, J.E. and Miller, J.L., 1985. Beyond crime seriousness: Fitting the punishment to the crime. *Journal of Quantitative Criminology*, *1*(1), pp.59-90.

Rossi, P. H., Waite, E., Bose, C. E., & Berk, R. E. (1974). The seriousness of crime: Normative structure and individual differences. *American Sociological Review*, 39, 224–237.

SafeLives., 2018 *About domestic abuse.* [Online] http://safelives.org.uk/policy-evidence/about-domestic-abuse [accessed 4th March 2019].

Saunders, D. G., 1992. A typology of men who batter women: three types derived from cluster analysis. *American Orthopsychiatry*, 62, 264–275.

Sebba, L., 1984. Crime seriousness and criminal intent. *Crime and Delinquency*, 30, 227–244.

Sellin, T., 1931. The basis of a crime index. *Am. Inst. Crim. L. & Criminology*, *22*, p.335.

Sellin, T., & Wolfgang, M., 1964. *The measurement of delinquency*. Montclair: Patterson Smith (Reprinted, 1978, with an introduction by S. Turner).

Sentencing Council., 2018. Sentencing guidelines for use in the Magistrate's Court. [Online] https://www.sentencingcouncil.org.uk/the-magistrates-court-sentencing-guidelines/ [accessed 30th December 2018].

Sexual Offences Act., 2003. *C80.* Available at: http://www.legislation.gov.uk/ukpga/2003/42/section/80 [accessed 3rd July 2019]

Sharp-Jeffs, N., 2015. *A review of research and policy on financial abuse within intimate partner relationships*. London, United Kingdom: London Metropolitan University.

Sharp-Jeffs, N., 2017. *Money matters: Research into the extent and nature of financial abuse within intimate relationships in the UK*. Co-operative Bank.

Shaw, D., 2016. Crime Severity Score measures 'relative harm' of crimes. [online] https://www.bbc.co.uk/news/uk-38157840 [accessed 4th March 2019].

Shepherd, J., 1990. Violent crime in Bristol: an accident and emergency department perspective. *The British Journal of Criminology*, *30*(3), pp.289-305.

Shepherd, J. P. 1998. Tackling violence: interagency procedures and injury surveillance are urgently needed. *British Medical Journal,* 316, 879-880.

Shepherd, J., & Lisles, C. 1998. Towards multi-agency violence prevention and victim support: an investigation of police-accident and emergency service liaison. *The British Journal of Criminology*, *38*(3), 351-370.

Sherman, L. W., 1992. *Policing Domestic Violence: Experiments and Dilemmas*, New York: Free Press.

Sherman, L. W. (1998). *Evidence-based policing*. Washington, DC: Police Foundation.

Sherman, L. W., 2007. The power few: experimental criminology and the reduction of harm. *Journal of Experimental Criminology*, *3*(4), pp.299-321.

Sherman, L.W., 2010. An introduction to experimental criminology. In *Handbook of quantitative criminology* (pp. 399-436). Springer, New York, NY.

Sherman, L.W., 2011. Al Capone, the Sword of Damocles, and the Police–Corrections Budget Ratio: Afterword to the Special Issue. *Criminology & Public Policy*, *10*(1), pp.195-206.

Sherman, L.W., 2013. The rise of evidence-based policing: Targeting, testing, and tracking. *Crime and justice*, *42*(1), pp.377-451.

Sherman, L.W., 2015. A tipping point for "totally evidenced policing" ten ideas for building an evidence-based police agency. *International criminal justice review*, *25*(1), pp.11-29.

Sherman, L.W., 2018. Evidence-based policing: Social organization of information for social control. In *Crime and social organization* (pp. 235-266). Routledge.

Sherman, L.W. and Berk, R.A., 1984. The specific deterrent effects of arrest for domestic assault. *American sociological review*, pp.261-272.

Sherman, L.W., Bland, M., House, P. and Strang, H., 2016. The Felonious Few vs. The Miscreant Many. *Cambridge: Cambridge Centre for Evidence Based Policing*.

Sherman, L.W. and Harris, H.M., 2015. Increased death rates of domestic violence victims from arresting vs. warning suspects in the Milwaukee Domestic Violence Experiment (MilDVE). *Journal of experimental criminology*, *11*(1), pp.1-20.

Sherman, L., Neyroud, P. W., & Neyroud, E., 2016. The Cambridge Crime Harm Index: measuring total harm from crime based on sentencing guidelines. *Policing: A Journal of Policy and Practice*, 10(3), 171-183.

Sherman, L.W., Schmidt, J.D. and Rogan, D.P., 1992. *Policing domestic violence: Experiments and dilemmas*. Free Press.

Sherman, L. W. & Strang, H, 1996. *Policing domestic violence: the problem-solving paradigm.* Paper presented at the Stockholm Conference on "Problem-Solving as Crime Prevention," Swedish National Council on Crime Prevention.

Smith, K., Flatley, J., Coleman, K., Osborne, S., Kaiza, P., & Roe, S. (2010). *Homicides, firearms offenses and intimate violence 2008/09* [Home Office Statistical Bulletin 01/10]. London, UK: Home Office.

Smith, C., 2016 'A Case Control Analysis of Offenders Issued with Domestic Violence Protection Orders (DVPOs) in Hertfordshire: A Retrospective and Prospective Study', M.St Thesis, Cambridge.

Sparrow, M.K., 2008. *The character of harms: Operational challenges in control*. Cambridge University Press.

Sparrow, M.K., 2011. Governing science. Harvard Kennedy School Program in Criminal Justice Policy and Management.

Stanley, N. and Humphreys, C., 2014. Multi-agency risk assessment and management for children and families experiencing domestic violence. *Children and youth services review*, *47*, pp.78-85.

Stark, E., 2007. *Coercive control: How men entrap women in everyday life*. New York, NY: Oxford University Press.

Stark, E., 2016. Policing Partner Abuse and the New Crime of Coercive Control in the United Kingdom. *Family & Intimate Partner Violence Quarterly*, *8*(4).

Steel, N., Blakeborough, L. and Nicholas, S., 2011. Supporting high-risk victims of domestic violence: a review of multi-agency risk assessment conferences (MARACs). *London: Home Office*.

Stoops, C., Bennett, L. and Vincent, N., 2010. Development and predictive ability of a behaviour-based typology of men who batter. *Journal of Family Violence*, *25*(3), pp.325-335.

Stout, K. D., 1993. Intimate femicide: A study of men who have killed their mates. *Journal of Offender Rehabilitation*, 19(3-4), 81-94.

Strang, H., 2012. Coalitions for a common purpose: managing relationships in experiments. *Journal of Experimental Criminology*, *8*(3), pp.211-225.

Strang, H., Sherman, L., Ariel, B., Chilton, S., Braddock, R., Rowlinson, T., Cornelius, N., Jarman, R. and Weinborn, C., 2017. Reducing the harm of intimate partner violence: Randomized controlled trial of the Hampshire Constabulary CARA Experiment. *Cambridge Journal of Evidence-Based Policing*, *1*(2-3), pp.160-173.

Strang, H., Sherman, L.W., Mayo-Wilson, E., Woods, D., Ariel, B. and Strang, H., 2013. Restorative Justice Conferencing (RJC) Using Face-to-Face Meetings of. *A Systematic Review. Campbell Systematic Reviews*, *12*.

Straus, M. A., 1990. Injury and frequency of assault and the "Representative sample fallacy" in measuring wife beating and child abuse. In M. A. Straus & R. J. Gelles edited with the assistance of C. Smith (Eds.), *Physical violence in American families: Risk factors and adaptations to violence in 8,145 families* (pp.75-91). New Brunswick, NJ: Transaction Publishers.

Stylianou, S., 2003. Measuring crime seriousness perceptions: What have we learned and what else do we want to know. *Journal of criminal justice*, *31*(1), pp.37-56.

Svalin, K., Mellgren, C., Torstensson Levander, M. and Levander, S., 2017. The inter-rater reliability of violence risk assessment tools used by police employees in Swedish police settings. *Nordisk Politiforskning; 1*, *4*.

The Independent, 2018. *A public health model is required to deal with knife crime – lives depend on it.* [Online] https://www.independent.co.uk/voices/editorials/sadiq-khan-london-knife-crime-glasgow-public-health-model-a8620421.html [accessed 4th March 2019].

Thornton, S., 2017. Police Attempts to Predict Domestic Murder and Serious Assaults: Is Early Warning Possible Yet? *Cambridge Journal of Evidence-Based Policing*, 1-17.

Tollenaar, N. and Van der Heijden, P.G.M., 2013. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(2), pp.565-584.

Turner, E., Medina, J and Brown, G. 2019. Dashing hopes? The predictive accuracy of domestic abuse risk assessment by the police, *The British Journal of Criminology*, azy074.

Tversky, A. and Kahneman, D., 1975. Judgment under uncertainty: Heuristics and biases. *In Utility, probability, and human decision making* (pp. 141-162). Springer Netherlands.

Tweed, R.G. and Dutton, D.G., 1998. A comparison of impulsive and instrumental subgroups of batterers. *Violence and victims*, *13*(3), pp.217-230.

Urwin, S., 2016. *Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers: An Assessment of Accuracy for the Durham Constabulary Model: Master Thesis*. University of Cambridge. Wolfson College.

Vigurs, C., Wire, J., Myhill, A., & Gough, D., 2016. *Police initial responses to domestic abuse*. London: College of Policing. Available at: http://whatworks.college.police.uk/Research/Documents/Police_initial_responses_domestic_abuse.pdf [accessed 4th March 2019]

Visher, C.A., Harrell, A., Newmark, L. and Yahner, J., 2008. Reducing intimate partner violence: an evaluation of a comprehensive justice system-community collaboration. *Criminology & Public Policy*, *7*(4), pp.495-523.

Von Hirsch, A. and Jareborg, N., 1991. Gauging criminal harm: a living-standard analysis. *Oxford J. Legal Stud.*, *11*, p.1.

Wagner, H. and Pease, K., 1978. On adding up scores of offence seriousness. *Brit. J. Criminology*, *18*, p.175.

Walby, S., 2005. Improving the statistics on violence against women. *Statistical Journal of the United Nations Economic Commission for Europe*, 22(3, 4), 193-216.

Walby, S., 2009. The Cost of Domestic Violence: Up-date 2009. Lancaster: Lancaster University

Walby, S., & Allen, J. (2004). Domestic violence, sexual assault and stalking: *Findings from the British Crime Survey*. Home Office.

Walker, J.R., 1997. *Estimates of the Costs of Crime in Australia 1996* (Vol. 72). Canberra: Australian Institute of Criminology.

Walker, L.E., 1979. *The battered woman*. New York: Harper & Row.

Walker, L.E., 1984. *The battered woman syndrome*. New York, NY: Springer.

Wallace, M., 2009. *Measuring crime in Canada: Introducing the crime severity index and improvements to the Uniform Crime Reporting Survey*. ProQuest.

Ward, T. and Stewart, C., 2003. Criminogenic needs and human needs: A theoretical model. *Psychology, Crime & Law*, *9*(2), pp.125-143.

Warr, M., 1989. What is the perceived seriousness of crimes. *Criminology*, 27, 795–814.

Watts, C. and Zimmerman, C., 2002. Violence against women: global scope and magnitude. *The lancet*, *359*(9313), pp.1232-1237.

Weinborn, C., Ariel, B., Sherman, L. W., & O'Dwyer, E., 2017. Hotspots vs. harmspots: Shifting the focus from counts to harm in the criminology of place. *Applied Geography*.

Weisburd, D., Groff, E.R. and Yang, S.M., 2012. *The criminology of place: Street segments and our understanding of the crime problem*. Oxford University Press.

Weisburd, D. and Neyroud, P., 2013. Police science: Toward a new paradigm. *Australasian policing*, *5*(2), p.13.

Weisz, A.N., Tolman, R.M. and Saunders, D.G., 2000. Assessing the risk of severe domestic violence: The importance of survivors' predictions. *Journal of interpersonal violence*, *15*(1), pp.75-90.

Wellford, C.F. and Wiatrowski, M., 1975. On the measurement of delinquency. *The Journal of Criminal Law and Criminology (1973-)*, *66*(2), pp.175-188.

Welsh, B.C., Farrington, D.P. and Gowar, B.R., 2015. Benefit-cost analysis of crime prevention programs. *Crime and justice*, *44*(1), pp.447-516.

Westmarland, N., Johnson, K. and McGlynn, C., 2017. Under the radar: The widespread use of 'out of court resolutions' in policing domestic violence and abuse in the United Kingdom. *The British Journal of Criminology*, *58*(1), pp.1-16.

Whinney, A., 2015. *A descriptive analysis of Multi-Agency Risk Assessment Conferences (MARACs) for reducing the future harm of domestic abuse in Suffolk.* University of Cambridge, Institute of Criminology.

Williams, A.E. and Ariel, B., 2012. The Bristol Integrated Offender Management Scheme: A pseudo-experimental test of desistance theory. *Policing: a journal of policy and practice*, *7*(2), pp.123-134.

Wilson, M., Daly, M., & Daniele, A., 1995. Familicide: The killing of spouse and children. *Aggressive Behavior*, 21(4), 275-291.

Wilson, M., & Daly, M., 1998. Lethal and nonlethal violence against wives and the evolutionary psychology of male sexual proprietariness. *Sage series on violence against women*, 9, 199-230.

Wire, J. & Myhill, A., 2018. *Piloting a new approach to domestic abuse frontline risk assessment*. Evaluation Report for the College of Policing [online] https://whatworks.college.police.uk/Research/Documents/DA_risk_assessment_pilot.pdf [accessed 4th March 2019].

Wolfgang, M., Figlio, R. and Sellin, T., 1972. Delinquency in a Birth Cohort. Chicago: Univ.

Wolfgang, M. E., Figlio, R. M., Tracy, P. E., & Singer, S. I., 1985. The national survey of crime severity. *Washington, DC: US Government Printing Office*, *204*.

Women's Aid., 2017. *How common is domestic abuse?* [Online] https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/how-common-is-domestic-abuse/ [accessed 4th March 2019].

Yang, M., Liu, Y. and Coid, J., 2010. Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism. *Ministry of Justice Research Series*, *6*(10).